Linear Regressions with Combined Data*

Xavier D'Haultfoeuille† Christophe Gaillac‡ Arnaud Maurel§

Abstract

We study linear regressions in a context where the outcome of interest and some of the covariates are observed in two different datasets that cannot be matched. Traditional approaches obtain point identification by relying, often implicitly, on exclusion restrictions. We show that without such restrictions, coefficients of interest can still be partially identified, with the sharp bounds taking a simple form. We obtain tighter bounds when variables observed in both datasets, but not included in the regression of interest, are available, even if these variables are not subject to specific restrictions. We develop computationally simple and asymptotically normal estimators of the bounds. Finally, we apply our methodology to estimate racial disparities in patent approval rates and to evaluate the effect of patience and risk-taking on educational performance.

Keywords: Data combination; best linear prediction; partial identification.

*First Version: December 6, 2024. We thank Pat Bayer, Christian Bontemps, Stephen Hansen, Marc Henry, Toru Kitagawa, Matt Masten, David Pacini, Daniel Wilhelm, and participants at seminars and conferences at the Encounters in Econometric Theory 2024, the Munich Econometrics Workshop 2024, 2024 ESEM (Rotterdam), LMU, Penn State University, the Aarhus Workshop in Econometrics V, the 35th $(EC)^2$ conference in Amsterdam, Toulouse School of Economics, University of Geneva (Statistics), University of Glasgow, University of Gothenburg, the Workshop on Optimal Transport in Econometrics (Collegio Carlo Alberto, 2025), and the 2025 IAAE Annual Conference. We thank Lavinia Kinne and Ludger Woessmann who kindly shared the PISA sample used in our paper. We also thank Yizhi Su and Haonan Ye for capable research assistance.

†CREST-ENSAE, xavier.dhaultfoeuille@ensae.fr.

‡University of Geneva, GSEM-IEE, christophe.gaillac@unige.ch.

§Duke University, NBER, CESifo and IZA, arnaud.maurel@duke.edu.

1 Introduction

It is often impossible to run the ideal regressions one would like to consider. A common reason for this is that the outcome Y and covariates X of interest are not observed in the same dataset. For instance, in many intergenerational studies (e.g., intergenerational income or wealth mobility), one cannot link parents' and children's outcomes. Even if the outcome and covariates of interest do appear in the same dataset, key control variables are often missing. For instance, when measuring the wage returns to education, one may wish to control for a measure of cognitive skills, but such measure may not be available in the main labor market dataset, even though it appears in another source.

To better explain our contribution in this context, we first detail our setup. We assume that X includes two sets of covariates: "outside" regressors X_o, which only appear in a separate dataset from that including the outcome Y, and "common" regressors X_c, which appear in both datasets. We also consider auxiliary variables, W_a, which researchers do not seek to include in the regression but that also appear in both datasets. These types of auxiliary variables are often available to empirical researchers. For instance, if a common variable is a proxy for a variable of interest X_o, or a so-called "bad control", it seems preferable to focus on the regression of Y on X_o, without controlling for that variable. We denote the set of common variables, included or not in the regression, by W, so that $W = (X_c', W_a')'$.

In this context, empirical researchers have traditionally relied on imputation methods. The most common, which corresponds to two-sample two-stage least squares (TSTSLS), consists in first predicting X_o by W_a in the "X_o dataset", and then using this prediction in the "Y dataset". One must recognize, however, that these approaches implicitly rely on exclusion restrictions and can therefore be sensitive to their violation. For instance, imputation based on TSTSLS requires the coefficient of W_a in the infeasible regression of Y on X and W_a to be 0. The goal of our paper is to study identification, estimation and inference on the regression coefficients without such exclusion restrictions.

In the absence of common variables, we obtain sharp bounds on each regression coefficient by applying the Frisch-Waugh-Lovell theorem together with the Cambanis-Simons-Stout inequality. Our contribution here is to show that this approach delivers sharp bounds, which is not obvious when X_o is multivariate. We then extend this result to account for common variables. Sharp bounds still

take a simple form in this case. Moreover, by leveraging the variation in W, one may be able to identify the sign of regression coefficients, and even obtain point identification in special cases. Importantly, we also show that it is possible to reject the exclusion restriction underlying the imputation based on TSTSLS.

Based on the identification results, we next turn to the estimation of the sharp bounds and inference on the regression coefficients. We propose simple plug-in estimators and establish their asymptotic normality. To do so, we build on results on L-statistics and on the statistical optimal transport literature. Our proof relies in particular on a refinement of the convergence rates in Fournier and Guillin (2015), which we obtain by extending a result of Boucheron and Thomas (2015). We also provide simple confidence intervals for the regression coefficients and establish their asymptotic validity. Simulation results indicate that our inference method performs well in finite samples, while being implementable at a modest computational cost.

Finally, we apply our methodology to two different contexts of data combination. In our first application, we revisit racial disparities in U.S. patent approval and show that conclusions about such disparities hinge on the validity of the exclusion restrictions underlying the TSTSLS estimation strategy. When relaxing these restrictions, our bounds are generally wide, pointing to the lack of robustness of the conclusions one would reach using TSTSLS results. In our second application, we evaluate the relationship between students' patience and risk-taking and educational performance across countries. In contrast to our first application, our bounds are informative and, for some of the specifications, exclude the TSTSLS estimates. Taken together, these applications highlight the limitations and, in some cases, misleading nature of the TSTSLS estimates in data combination environments. The partial identification approach we propose in this paper provides a transparent and tractable way to assess the sensitivity of empirical conclusions to the exclusion restrictions underlying the TSTSLS estimates.

Related literature. To our knowledge, the first paper that considered our problem is Pacini (2019). We extend his work in three important dimensions. First, we study the case where some of the common variables are not used as common regressors ($W \neq X_c$). We expect this to be prevalent in practice and we show that it can drastically reduce the identified sets. Second, we show that his bounds are not sharp when X_o is multivariate, and that the difference with the sharp

bounds can be substantial. Finally and importantly, we consider estimation and inference. Hwang (2025) also relates closely to our work. While she maintains the restriction that $W = X_c$, she also considers the case where some regressors are only available in the Y dataset, which we do not study here.[1] A third related study is Fan et al. (2025). This paper complements ours by studying identification in a more general setup. In particular, they derive sharper bounds than Hwang (2025) in cases where some regressors are observed only in the Y dataset. Their analysis, however, is limited to identification, whereas estimation, inference and empirical applicability are central to our paper.

Our paper is also related to our own previous work (D'Haultfœuille et al., 2025), in which we consider a similar data combination environment. There are, however, important differences between the two. First, we did not consider previously how auxiliary variables affect identification. Second, we imposed a partially linear model, namely $E[Y|X] = X_o'\beta_o + f(X_c)$. This leads to potentially tighter bounds, but one may be reluctant to improve bounds using such restrictions. Third, for estimation we had to focus on the case for which X_c had finite support, whereas no such assumption is necessary here. Finally, from a technical viewpoint, the restriction on the conditional expectation implies that we relied on entirely different optimal transport results, both for identification and inference.

At a broader level, our paper belongs to a very active literature on data combination problems in econometrics and statistics. See, in particular, Ridder and Moffitt (2007) for a survey of this literature and contributions by Fan et al. (2014), Fan et al. (2016), Buchinsky et al. (2022), Bontemps et al. (2025), Meango et al. (2025) and, in the context of experimental data under a surrogacy assumption, Athey et al. (2020), Athey et al. (2024), and Rambachan et al. (2024). Several of these papers impose restrictions that entail point identification. Following the seminal contribution of Cross and Manski (2002) and subsequent article by Molinari and Peski (2006), our aim is to obtain bounds on parameters of interest under weak restrictions. An important distinction between our work and these last two papers is that we consider different parameters: the best linear parameter in our case versus conditional expectation in theirs. Also, they do not consider the use

[1]There are still other data combination cases that we do not consider here. Kitawaga and Sawada (2023) consider a setup where one observes (Y, X_1, X_c) in one dataset and (Y, X_2, X_c) in another. Yet another possibility, considered by Moon (2024) when X_1, X_2, X_c has finite support, is to observe (Y, X_c), (X_1, X_c) and (X_2, X_c) separately.

of auxiliary variables (W_a).

From a technical viewpoint, our first identification result can be seen as an extension of the Cambanis-Simons-Stout inequality, see Cambanis et al. (1976) and, e.g., Fan et al. (2014, 2016) for an application to data combination problems. Our asymptotic normality result relates to the asymptotic normality of the so-called Wasserstein-2 distance of empirical measures, recently studied in the statistical literature (see, e.g. Del Barrio et al., 2019; Berthet et al., 2020). Notably, up to a mild strengthening of moment conditions (from order 4 to $4+\varepsilon$ for some $\varepsilon > 0$), our result implies asymptotic normality of the Wasserstein-2 distance under weaker conditions than those in Berthet et al. (2020).

Finally, our paper also speaks to a large and growing empirical literature that deals with data combination problems similar to the one considered here. One important example is voting: given the anonymity of ballots, researchers typically regress average votes on average voter characteristics (e.g., income, hours watching Fox News per week) at the county level (see, e.g., Martin and Yurukoglu, 2017). This approach implicitly relies on a TSTSLS strategy, where counties play the role of W_a, thereby imposing an exclusion restriction. Another leading example is intergenerational income mobility, which often faces the unavailability of linked income data across generations and similarly relies on exclusion restrictions (see, e.g., Santavirta and Stuhler, 2024, for a survey). Data combination issues are also pervasive in consumption research, where income and consumption are often measured in separate datasets (see in particular Crossley et al., 2022, who discuss another imputation strategy than that based on TSTSLS). Similar data combination problems frequently arise in various other subfields, including the economics of education and returns to skill estimation (Piatek and Pinger, 2016; Garcia et al., 2020; Hanushek et al., 2022), health (Manski, 2018) and labor (Athey et al., 2020). Finally, gaps in science and innovation by race or gender provide another relevant example, as illustrated in our first application below.

The methods we devise in this paper are broadly applicable in these different contexts, allowing empirical researchers to relax the exclusion restrictions that are typically maintained to achieve point identification. By applying our method to racial disparities in patent approval (Dossi, 2024) and the effect of preferences on skill differences (Hanushek et al., 2022), our paper also adds to the empirical literature on these questions.

Outline. Section 2 introduces the setup and discusses three broad cases for which our analysis is relevant. Section 3 presents our identification results. Section 4 develops estimators of the sharp bounds, establishes their asymptotic normality and develops inference on the regression coefficients. Section 5 examines the finite sample properties of our estimators and confidence intervals through Monte Carlo simulations. We provide in Section 6 two applications, to racial disparities in patent approval and the effect of preferences on skill differences. Finally, Section 7 concludes. The appendix includes in particular a discussion of the sharpness of the bounds of Pacini (2019) and gathers all the proofs of our identification results; the proofs of our inference results appear in the online appendix. Finally, our method can be implemented using our companion R package, `RegCombinBLP`.[2]

2 Set-up and motivation

2.1 Set-up

We seek to identify the best linear predictor $EL(Y|X)$ of Y by $X \in \mathbb{R}^p$, with $X = (X_o', X_c')'$. To this end, we assume to have access to two separate datasets that cannot be matched. The first one includes (Y, W'), whereas the second one includes (X_o', W'); here $W = (W_a', X_c') \in \mathbb{R}^q$. We call X_o the "outside regressors", X_c the "common regressors", W the "common variables" and W_a the "auxiliary variables". The latter are variables that the researcher does not want to include in the regression of interest, but that may still help for identification since they are included in both datasets. Importantly, they should not be seen as instruments, in the sense that we do not impose below any restrictions on them.

In order for the best linear prediction to be well-defined, we maintain the following assumption hereafter:

Assumption 1 *$E(Y^2 + \|X_o\|^2 + \|W\|^2) < \infty$ and $E(XX')$ and $E(WW')$ are nonsingular.*

Let $b^0 = (b^{0,1}, ..., b^{0,p}) \in \mathbb{R}^p$ be such that $EL(Y|X) = X'b^0$. Usually, researchers are interested in specific components of b^0, rather than in the whole vector b^0. Therefore, in the following we seek to (partially) identify and estimate $b_d := d'b^0$,

[2]This package is available on GitHub at https://github.com/cgaillac/RegCombinBLP with a user-friendly guide on how to use it.

for some $d \in \mathbb{R}^p$. For instance, if we focus on $b^{0,2}$, the second component of b^0, we let $d = (0, 1, 0, ..., 0)$.

2.2 Motivation

Our setup includes at least three cases of broad interest.

Proxies for the covariate of interest. In this case, we are interested in the relationship between a covariate of interest X_o and Y. However, we do not observe X_o in the Y dataset, but only proxies W_a of X_o. These proxies are also observed in the X_o dataset. This type of situation arises very frequently in empirical microeconomics. A standard strategy in this case is to use two-sample two-stage least squares (TSTSLS). Namely, one first regresses X_o on W_a in the X_o dataset. Then, we regress Y on the predicted X_o in the Y dataset. Importantly though, this strategy implicitly relies on the following exclusion restriction:

$$EL(Y|X_o, W_a) = EL(Y|X_o). \tag{1}$$

This assumption is often restrictive. In our first application below, for instance, Y corresponds to patent approval, X_o is the vector of race dummies and W_a is the vector of applicants' last names. Given that patent reviewers always observe last names but typically do not observe race directly, (1) seems unlikely to hold. The method we develop in this paper will allow us to (partially) identify $EL(Y|X_o)$ without imposing (1).

Missing controls. In this case, we are interested in recovering the effect of X_c on Y using data from a first dataset. However, one or several key control variables (X_o) are missing from this dataset. Our setup applies to situations where the control variables are observed in a second dataset, together with X_c. In this sense, our framework complements a growing literature that investigates how credible unconfoundedness is, by allowing researchers to rely on unconfoundedness in a broad range of data combination environments (see, e.g., Altonji et al., 2005; Oster, 2019; Diegert et al., 2022).

As above, researchers in this context may also have access to auxiliary variables, W_a, which are not included as covariates in the main regression, e.g., because they would be "bad controls". As shown below, these variables may still carry informational content regarding the regression coefficients of interest.

Mediation analysis. In this case, we are interested in the effect of X_o on a given outcome Y. As above, we do not observe X_o and Y in the same dataset. A possible and frequent reason is that Y is a long-run outcome, which is not observed in the data including X_o. On the other hand, both datasets may include other outcomes W_a, such as short-run outcomes.

To identify in this environment the causal effect of X_o on Y (which is b^0 under suitable randomization conditions on X_o), a common strategy is to rely on a surrogacy assumption (see, e.g., Prentice, 1989). In our setup, this corresponds to

$$EL(Y|X_o, W_a) = EL(Y|W_a). \tag{2}$$

In other words, one assumes that the effect of X_o on Y is entirely mediated by W_a.[3] However, Condition (2) typically is a strong restriction. For instance, it is reasonable to assume that long-run earnings (Y) depend on human capital, even conditional on short-run earnings (W_a). Then, if job training (X_o) affects human capital, (2) will fail to hold in general. Our results below imply that one can still obtain simple, sharp bounds on b^0, without relying on (2).

3 Identification

Before presenting our identification results, we introduce additional notation. We denote by $\mathcal{B}_d$ the identified set of b_d and let $\overline{b}_d$ and $\underline{b}_d$ be their sharp upper and lower bounds, namely

$$\overline{b}_d = \sup\{d'b : \ b \in \mathcal{B}\}, \quad \underline{b}_d = \inf\{d'b : \ b \in \mathcal{B}\},$$

where $\mathcal{B}$ is the identified set of b^0. We focus in the following solely on $\overline{b}_d$, which is without loss of generality since $\underline{b}_d = -\overline{b}_{-d}$.

For any random variables A and B, we let F_A denote the cumulative distribution function (cdf) of A, f_A its density, and $F_{A|B}$ the cdf of A given B. We also let $F_A^{-1}(t) := \inf\{x : F_A(x) \geq t\}$ denote the quantile function of A; we denote similarly by $F_{A|B}^{-1}$ the quantile function of A given B. We let $\text{Supp}(A)$ (resp. $\text{Supp}(A|B)$) denote the support of the probability distribution of A (resp., of A given B). For any vector v, we let v_k denote its k-th element and v_{-k} the vector

[3] One may also include additional covariates X_c observed in both datasets, in which case (2) becomes $EL(Y|X, W_a) = EL(Y|X_c, W_a)$.

obtained by removing v_k from v. We also let $e_{k,r}$ denote the k-th canonical vector of $\mathbb{R}^r$. For any set S, we let $|S|$ denote its cardinality. Finally, we denote by $\mathcal{U}[0,1]$ the uniform distribution over $[0,1]$ and by $\mathcal{N}(\mu,\Sigma)$ the multivariate normal distribution with mean μ and covariance matrix Σ.

3.1 No common variables

We first consider a case without nontrivial common variable ($W = X_c = 1$), so that $X = (X_o', 1)' \in \mathbb{R}^p$. Our main result shows that $\mathcal{B}$ is convex and compact, and characterizes $\bar{b}_d$ for any $d \in \mathbb{R}^p \backslash \{0\}$. Below, we introduce the variable η_d as follows. First, let $(d_2, ..., d_p)$ be $(p-1)$ vectors in $\mathbb{R}^p$ such that $(d, d_2, ..., d_p)$ forms a basis of $\mathbb{R}^p$. Let M denote the corresponding matrix and let $T = M^{-1}X$. Then, let

$$\eta_d := T_1 - EL[T_1|T_{-1}].$$

In words, η_d is the residual of the (population) regression of T_1 on T_{-1}. Note that η_d does not depend on which exact vectors $(d_2, ..., d_p)$ are chosen. Also, if $d = e_{k,p}$, η_d is simply the residual of the regression of X_k on X_{-k}. Finally, if $p = 2$ and $d = (d_1, 0)'$, $\eta_d = (X_o - E(X_o))/d_1$.

Theorem 1 *Suppose that Assumption 1 holds and $W = X_c = 1$. Then $\mathcal{B}$ is convex, compact, and satisfies $\mathcal{B} \subseteq \mathcal{E}$, with*

$$\mathcal{E} := \{b \in \mathbb{R}^p : \ E[Y] = E[X'b], \ V(Y) \geq V(X'b)\}.$$

Also, letting $U \sim \mathcal{U}[0,1]$, we have, for any $d \in \mathbb{R}^p \backslash \{0\}$, $\mathcal{B}_d = [\underline{b}_d, \bar{b}_d]$, with

$$\bar{b}_d = E\left[F^{-1}_{d'E(XX')^{-1}X}(U) F_Y^{-1}(U)\right] \tag{3}$$

$$= \frac{E[F^{-1}_{\eta_d}(U) F_Y^{-1}(U)]}{E(\eta_d^2)}. \tag{4}$$

Finally, $\bar{b}_d > 0$ as long as $V(Y) > 0$.

The first part of the theorem states that $\mathcal{B}$ is a convex, compact set included in the ellipsoid $\mathcal{E}$. Also, $(0, ..., 0, E[Y])' \in \mathcal{B}$: in the absence of common variables, we can always rationalize that Y and X are independent. Since the identified set $\mathcal{B}$ is non-empty, closed, and convex, $\bar{b}_d$ is equal to the so-called support function of $\mathcal{B}$. As a result, the knowledge of $\bar{b}_d$ for all $d \in \mathbb{R}^p \backslash \{0\}$ characterizes $\mathcal{B}$.

In the case of a single regressor (and the intercept) and $d = (1, 0)'$, Equation (4) reduces to

$$\overline{b}_d = \frac{E\left[(F_{X_o}^{-1}(U) - E(X_o))F_Y^{-1}(U)\right]}{V(X_o)}. \tag{5}$$

On the other hand, the true coefficient satisfies $b_d = b^{0,1} = E[(X_o - E(X_o))Y]/V(X_o)$. Thus, (5) indicates that the sharp upper bound on the unknown term $E[X_oY]$ is $E[F_{X_o}^{-1}(U)F_Y^{-1}(U)]$. This is well-known, and corrresponds to the so-called Cambanis-Simons-Stout inequality (see Cambanis et al., 1976). The logic is that (i) $F_{X_o}^{-1}(U)$ and $F_Y^{-1}(U)$ are distributed as X_o and Y, since U is uniformly distributed, and (ii) these two variables exhibit maximal positive dependence. The exact meaning of (ii) is that the copula of $F_{X_o}^{-1}(U)$ and $F_Y^{-1}(U)$ corresponds to the Fréchet-Hoeffding upper bound.

With multiple regressors, (4) cannot be directly deduced from the Cambanis-Simons-Stout inequality. To get some intuition on (4), suppose that $d = e_{1,p}$. Then, η_d is the residual of the linear regression of X_1 on X_{-1}. If we observed (Y, X), the coefficient of X_1 in the best linear prediction of Y by X would be $E[\eta_d Y]/E(\eta_d^2)$, by the Frisch-Waugh-Lovell theorem. Now, if we only know the marginal distributions of η_d and Y, the numerator in (4) is simply the upper bound of $E[\eta_d Y]$. That the sharp upper bound $\overline{b}_d$ satisfies (4) is not obvious, however, because we also know the distribution of X_{-1} conditional on η_d, in addition to the marginal distribution of η_d. This could, in principle, lead to $\overline{b}_d < E[F_{\eta_d}^{-1}(U)F_Y^{-1}(U)]/E(\eta_d^2)$. Theorem 1 shows that this is not the case: the conditional distribution of X_{-1} does not carry any additional information about $E[\eta_d Y]$. Although this can be deduced from Lemma 3.3 in Delon et al. (2023), we propose an alternative proof, which has the advantage of being constructive.

Pacini (2019) also obtains bounds on b_d, see his Theorem 1. However, it turns out that when X is multivariate, his bound is only an outer bound rather than the sharp bound $\overline{b}_d$ on b_d. In Appendix A, we detail why this is the case, and provide an illustration showing that the sharp bounds given by Theorem 1 above can in practice be substantially tighter than Pacini's bounds.

3.2 Common variables

3.2.1 Main result

Let us now turn to the situation where some covariates are observed in both datasets. We define as above η_d, with the sole difference that now $X = (X_o', X_c')'$. Next, let δ_d and ν_d be such that $EL(\eta_d|W) = W'\delta_d$ and $\nu_d := \eta_d - W'\delta_d$. Define δ_Y and ν_Y similarly, with Y in place of η_d. The following theorem is the counterpart of Theorem 1 with common variables.

Theorem 2 *Suppose that Assumption 1 holds. Then $\mathcal{B}$ is convex, compact, and for any $d \in \mathbb{R}^p \backslash \{0\}$, $\mathcal{B}_d = [\underline{b}_d, \overline{b}_d]$, with*

$$\overline{b}_d = \frac{1}{E(\eta_d^2)} \left\{ \delta_d' E(WW')\delta_Y + E\left[F^{-1}_{\nu_d|W}(U|W) F^{-1}_{\nu_Y|W}(U|W) \right] \right\}, \tag{6}$$

where $U|W \sim \mathcal{U}[0,1]$. Moreover, for any function g, $\overline{b}_d \leq \overline{b}_d^g$, with

$$\overline{b}_d^g := \frac{1}{E(\eta_d^2)} \left\{ \delta_d' E(WW')\delta_Y + E[F^{-1}_{\nu_d|g(W)}(U|g(W)) F^{-1}_{\nu_Y|g(W)}(U|g(W))] \right\}, \tag{7}$$

with equality if $\nu_Y \perp\!\!\!\perp W|g(W)$ and $\nu_d \perp\!\!\!\perp W|g(W)$.

Essentially, the first part of the theorem follows by first applying Theorem 1 conditional on W and then integrating over W. The second part exploits Theorem 1 but conditioning on $g(W)$ instead of W. The sharp bound $\overline{b}_d$ has a simple expression, but it involves the conditional quantile functions $F^{-1}_{\nu_d|W}$ and $F^{-1}_{\nu_Y|W}$. Thus, estimating this sharp bound involves estimating these two nonparametric functions, which could be cumbersome in practice. On the other hand, when $g(W)$ has a finite support, the outer bound $\overline{b}_d^g$ is elementary to estimate and does not suffer from any curse of dimensionality. Moreover, this bound is actually sharp when $\nu_Y \perp\!\!\!\perp W|g(W)$ and $\nu_d \perp\!\!\!\perp W|g(W)$, as is the case for instance with $g(W) = 1$, if Y and η_d follow a linear location model in W.

How do common regressors affect identification? Even without auxiliary variables W_a, the identified interval on the coefficients of X_o may exclude 0 in the presence of common regressors, implying that the sign of these coefficients is identified. To see this, suppose that $\dim(X_o) = 1$, $X_o = f_1(X_c) + \zeta_o$, $Y = g_1(X_c) + \zeta_Y$ and $\zeta_o|X_c \sim \mathcal{N}(0, \sigma_o^2)$, $\zeta_Y|X_c \sim \mathcal{N}(0, \sigma_Y^2)$. Let also $f(X_c) := f_1(X_c) - EL(f_1(X_c)|X_c)$ and $g(X_c) := g_1(X_c) - EL(g_1(X_c)|X_c)$. Using Equation (6), the

fact that by construction $\delta_d = 0$, and the normality of ζ_o and ζ_Y, we obtain that the bounds on $b^{0,1}$ satisfy

$$\begin{aligned}\overline{b}_{e_1} =& \frac{E\left[f(X_c)g(X_c)\right] + \sigma_o\sigma_Y}{E(\eta_{e_1}^2)},\\ \underline{b}_{e_1} =& \frac{E\left[f(X_c)g(X_c)\right] - \sigma_o\sigma_Y}{E(\eta_{e_1}^2)},\end{aligned}$$

where $\eta_{e_1} = f(X_c) + \zeta_o$. In particular, if $|E\left[f(X_c)g(X_c)\right]| > \sigma_o\sigma_Y$, 0 is excluded from the identified set of $b^{0,1}$. This occurs when X_o and Y strongly depend on X_c in a nonlinear way, so that $E\left[f(X_c)g(X_c)\right]$ dominates the contribution from independent terms (namely, $\sigma_o\sigma_Y$). In the extreme case where X_o and Y are deterministic functions of X_c, so that $\sigma_o = \sigma_Y = 0$, we obtain point identification.

That said, the identified interval on the coefficients of X_o may widen when including common covariates. Even if observing X_c in both dataset does increase the information on the joint distribution of (Y, X_o), the parameter we consider also changes. In particular, the denominator $E[\eta_d^2]$ in (6) may substantially decrease, if the R^2 of the linear regression of X_o on X_c is large. To illustrate this, suppose that $(X_o, X_c')' \sim \mathcal{N}(0, \Sigma_o)$ and $(Y, X_c') \sim \mathcal{N}(0, \Sigma_Y)$, with

$$\Sigma_o = \begin{pmatrix} 1 & \rho_o \\ \rho_o & 1 \end{pmatrix} \quad \text{and } \Sigma_Y = \begin{pmatrix} 1 & \rho_Y \\ \rho_Y & 1 \end{pmatrix}. \tag{8}$$

Then, some algebra shows that without observing X_c, $[\underline{b}_{e_1}, \overline{b}_{e_1}] = [-1, 1]$. With X_c, on the other hand,

$$[\underline{b}_{e_1}, \overline{b}_{e_1}] = \left[-\sqrt{\frac{1-\rho_Y^2}{1-\rho_o^2}},\ \sqrt{\frac{1-\rho_Y^2}{1-\rho_o^2}}\right].$$

Thus, the interval $[\underline{b}_{e_1}, \overline{b}_{e_1}]$ shrinks if $|\rho_Y| > |\rho_o|$ but widens otherwise.

Finally, we may also identify the sign of components of b_c, the regression coefficient of X_c. In fact, b_c may even be point identified: if X_c and X_o are uncorrelated, b_c is simply the coefficient of the regression of Y on X_c.

The role of auxiliary variables. By observing auxiliary variables W_a that are not in the regression of interest, we increase the available information without modifying the parameter of interest. Then, the interval $[\underline{b}_{e_1}, \overline{b}_{e_1}]$ always shrinks (at least weakly so). This may lead to excluding 0 from $\mathcal{B}$ even without common variables X_c, a case that occurs whenever

$$\delta_d' E(WW')\delta_Y + E\left[F^{-1}_{\nu_d|W}(U|W)F^{-1}_{\nu_Y|W}(U|W)\right] < 0 \tag{9}$$

for some $d \in \mathbb{R}^p$. Intuitively, (9) requires enough dependence between Y and W_a and between X_o and W_a. For instance, if $(W, X_o) \sim \mathcal{N}(0, \Sigma_o)$ and $(W, Y) \sim \mathcal{N}(0, \Sigma_Y)$, with Σ_o and Σ_Y as in (8), we obtain

$$[\underline{b}_{e_1}, \overline{b}_{e_1}] = \left[\rho_o \rho_Y - \sqrt{(1-\rho_o^2)(1-\rho_Y^2)},\ \rho_o \rho_Y + \sqrt{(1-\rho_o^2)(1-\rho_Y^2)}\right].$$

Then, $0 \notin [\underline{b}_{e_1}, \overline{b}_{e_1}]$ if and only if $\rho_o^2 \rho_Y^2 > (1-\rho_o^2)(1-\rho_Y^2)$. This holds when W_a is sufficiently correlated with X_o and Y. For instance, when $\rho_o = \rho_Y$, this occurs if and only if W_a explains more than half of the variance of X_o ($\rho_o^2 > 1/2$).

A leading case with auxiliary variables is the case of surrogates. Recall that in this case, X_o corresponds to the treatment variable, Y is a long-run outcome while W_a denotes short-run outcomes (surrogate variable). Then, Theorem 2 yields two sets of bounds, sharp and outer, on the effect of X_o on Y, without imposing a surrogacy assumption.

Link with TSTSLS. Recall that the TSTSLS estimand identifies b^0 if the coefficient of W_a in the "long" regression of Y on (X, W_a) is 0. Now, the discussion above ("How do common regressors affect identification?") shows that if one views W_a as a common regressor, 0 may not belong to the identified set of the regression coefficient of W_a. This implies that the exclusion restriction underlying the TSTSLS estimand can actually be rejected by the data. As a simple example, suppose that X_o and W_a are not correlated. Then, the coefficient of W_a in the "long" regression is equal to the coefficient of W_a in the "short" regression of Y on W_a, and this coefficient may not be 0. Beyond this particular case, the TSTSLS estimand for the coefficient $b^{0,k}$ may not belong to the sharp identified set $[\underline{b}_{e_k}, \overline{b}_{e_k}]$, something we illustrate in our second application below.

3.2.2 Testing and weakening the common population assumption

We have maintained thus far that the two samples at hand are drawn from the same population. While this is a standard assumption in the data combination literature, it is important to consider the extent to which this can be relaxed. To this end, let us introduce the binary variable D, with $D = 1$ (resp. $D = 0$) if we consider the Y dataset (resp. the X_o dataset). Then, our setup implies that we only observe the distributions of $(W, Y)|D = 1$ and $(W, X_o)|D = 0$, assuming that $D \perp\!\!\!\perp (W, X_o, Y)$. With common variables, this condition can be tested, since

it implies $F_{W|D=1} = F_{W|D=0}$. If this implication is rejected, we can weaken the independence assumption by assuming instead that

$$(X_o, Y) \perp\!\!\!\perp D|W, \quad p := P(D = 1) \text{ is known.} \tag{10}$$

In words, the first condition imposes that conditional on W, the two datasets are drawn from the same population, while the two populations corresponding to $D = 0$ and $D = 1$ may differ in their marginal distributions of W. The second condition in (10) implies that the joint distribution of (D, W), and thus the "propensity score" $p(W) := P(D = 1|W)$, can be retrieved from the knowledge of the distributions of $W|D = 0$ and $W|D = 1$.

If (10) holds, the sharp upper bound $\overline{b}_d$ can be obtained by reasoning as in Theorem 2, using an inverse probability weighting scheme. Specifically, to identify $\delta_Y = E[WW']^{-1}E[WY]$ (and then ν_Y), we cannot directly regress Y on W conditional on $D = 1$. Yet, we can recover it by considering instead a weighted regression, as

$$\delta_Y = E\left[\frac{DWW'}{p(W)}\right]^{-1} E\left[\frac{DWY}{p(W)}\right].$$

We can identify δ_d (and then ν_d) similarly, using the weights $(1 - D)/(1 - p(W))$. Then, Equation (6) is replaced by:

$$\overline{b}_d = \frac{1}{E\left[\frac{(1-D)\eta_d^2}{1-p(W)}\right]} \left\{\delta_d' E(WW')\delta_Y + E\left[F^{-1}_{\nu_d|W,D=0}(U|W)F^{-1}_{\nu_Y|W,D=1}(U|W)\right]\right\}.$$

Another point to note is that if the two populations differ, the parameter of interest may correspond to one of the two populations only. For instance, one may consider, instead of $EL(Y|X)$, $EL(Y|X, D = 1)$. In this case, δ_Y is given by $E[WW'|D = 1]^{-1}E[WY|D = 1]$ and is thus obtained by an unweighted regression, whereas δ_d (and then ν_d) is obtained by regressing η_d on W with weights $p(W)/(1 - p(W))$. The upper bound $\overline{b}_d$ becomes

$$\overline{b}_d = \frac{E(D)\left\{\delta_d' E(WW'|D = 1)\delta_Y + E\left[F^{-1}_{\nu_d|W,D=0}(U|W)F^{-1}_{\nu_Y|W,D=1}(U|W)|D = 1\right]\right\}}{E\left[(1 - D)\eta_d^2 p(W)/(1 - p(W))\right]}.$$

Finally, another practically relevant situation is one in which one sample is drawn from a subpopulation of the population from which the other sample is drawn. Then, we identify instead (for instance) the distribution of (Y, W) given $D = 1$ and the distribution of (X, W). In this case and if we focus as above on $EL(Y|X, D = 1)$, we obtain a similar upper bound on $\overline{b}_d$ as above, with just a few differences.

First, δ_d (and then ν_d) is obtained by regressing η_d on W with weights $p(W)$. Second, we now have

$$\bar{b}_d = \frac{E(D)}{E\left[p(W)\eta_d^2\right]}\left\{\delta_d' E(WW'|D=1)\delta_Y + E\left[F^{-1}_{\nu_d|W}(U|W)F^{-1}_{\nu_Y|W,D=1}(U|W)|D=1\right]\right\}. \tag{11}$$

Note that in this case and the one before, we do not require the joint independence condition in (10) but only $X_o \perp\!\!\!\perp D|W$.

3.2.3 Auxiliary, non-common variables

In practice, one may have access to auxiliary variables that appear in the dataset of Y only, or in the dataset of X_o only. For instance, suppose we identify the distributions of (W, Y, Z) from one dataset and that of (W, X_o) from the other. The following proposition shows that, for identification purposes, knowing the conditional distribution of $Z|W, Y$ provides no additional information. Hereafter, we let $\mathcal{B}_Z$ denote the identified set of b^0 when observing some auxiliary non-common variables Z.

Proposition 1 *Suppose that Assumption 1 holds. Then $\mathcal{B}_Z = \mathcal{B}$.*

A similar result clearly holds if we consider instead a variable that appears only in the dataset of X_o. The bottom line is that, among variables not included in the regression, only those that are common across the two datasets are relevant for identification.

4 Estimation and inference

4.1 Estimation of $\bar{b}_d$

4.1.1 No common variables

Consider first the simplest situation where we only observe two independent samples, $\mathcal{S}_1 := (Y_i)_{i=1,\dots,n}$ and $\mathcal{S}_2 := (X_j)_{j=1,\dots m}$. Let $\widehat{\eta}_{dj}$ denote j's residual in the sample regression of T_1 on T_{-1} (recall the definition of T at the beginning of Section 3.1). To ease notation, we let hereafter $F := F_Y$ and $G := F_{\eta_d}$, and let F_n and $\widehat{G}_m$ denote the empirical cdfs of $(Y_i)_{i=1,\dots,n}$ and $(\widehat{\eta}_{dj})_{j=1,\dots,m}$, respectively. From Theorem 1, we have $\bar{b}_d = \int_0^1 F^{-1}(t)G^{-1}(t)dt/E(\eta_d^2)$. Then, we consider the

plug-in estimator of $\overline{b}_d$:

$$\widehat{\overline{b}}_d = \frac{\int_0^1 F_n^{-1}(t)\widehat{G}_m^{-1}(t)dt}{\widehat{E}(\widehat{\eta}_d^2)},$$

where $\widehat{E}(\widehat{\eta}_d^2)$ denotes the empirical variance of $(\widehat{\eta}_{dj})_{j=1,...,m}$. Remark that when $m = n$, we simply have, denoting by $Y_{(i)}$ the i-th order statistic of $(Y_i)_{i=1,...,n}$ (similarly for $\widehat{\eta}_{d(i)}$):

$$\int_0^1 F_n^{-1}(t)\widehat{G}_m^{-1}(t)dt = \frac{1}{n}\sum_{i=1}^n Y_{(i)}\widehat{\eta}_{d(i)}.$$

Otherwise, we can still compute the numerator of $\widehat{\overline{b}}_d$ at low cost. To see this, note that for any real-valued variables U_1, U_2 with finite second moments and cdfs F_1, F_2,

$$\int_0^1 F_1^{-1}(t)F_2^{-1}(t)dt = \frac{1}{2}\left[E[U_1^2] + E[U_2^2] - W_2^2(F_1, F_2)\right], \tag{12}$$

where W_2 is the Wasserstein-2 distance, $W_2(F_1, F_2) := (\int (F_2^{-1}(t) - F_1^{-1}(t))^2 dt)^{1/2}$. For variables with support size of n and m respectively, as is the case here, we can then compute $W_2(F_1, F_2)$ with algorithms of complexity $O(m + n)$, see, e.g., Rubner et al. (2000).

4.1.2 Common variables

Let us now consider the case where common variables are observed. Specifically, we now assume to observe $\mathcal{S}_1 := \{(Y_1, W_1^{(1)}), ..., (Y_n, W_n^{(1)})\}$ and $\mathcal{S}_2 := \{(X_{o1}, W_1^{(2)}), ..., (X_{om}, W_m^{(2)})\}$, where $W_i^{(1)}$ and $W_j^{(2)}$ are both distributed as W. We add the exponents (ℓ) to indicate that $W^{(\ell)} \in \mathcal{S}_\ell$. Recall from Theorem 2 that $\overline{b}_d$ involves the nonparametric functions $F^{-1}_{\nu_d|W}$ and $F^{-1}_{\nu_Y|W}$. To avoid their estimation, we consider instead the outer bound $\overline{b}_d^g$ for a function g taking finitely many values $(g_1, ..., g_K)$. Then,

$$\overline{b}_d^g = \frac{1}{E(\eta_d^2)}\left\{\delta_d' E(WW')\delta_Y + \sum_{k=1}^K p_k F_{|k}^{-1}(U)G_{|k}^{-1}(U)\right\},$$

where $p_k := P(g(W) = g_k)$, $F_{|k} := F_{\nu_Y|g(W)}(.|g_k)$ and $G_{|k} := F_{\nu_d|g(W)}(.|g_k)$. Again, we consider a plug-in estimator of $\overline{b}_d^g$:

$$\widehat{\overline{b}}_d^g = \frac{1}{\widehat{E}(\widehat{\eta}_d^2)}\left\{\widehat{\delta}_d'\widehat{E}(WW')\widehat{\delta}_Y + \sum_{k=1}^K \widehat{p}_k \int_0^1 \widehat{F}_{|k}^{-1}(u)\widehat{G}_{|k}^{-1}(u)du\right\},$$

where $\widehat{F}_{|k}$ (resp. $\widehat{G}_{|k}$) is the empirical cdf of $\widehat{\nu}_Y$ (resp. $\widehat{\nu}_d$) on the subsample of $\mathcal{S}_1$ satisfying $g(W_i^{(1)}) = g_k$ (resp., the subsample of $\mathcal{S}_2$ satisfying $g(W_j^{(2)}) = g_k$). The

estimators $\widehat{E}(WW')$ and $\widehat{p}_k$ are simply obtained by combining the two samples, e.g.,

$$\widehat{E}(WW') := \frac{1}{m+n}\left[\sum_{i=1}^{n} W_i^{(1)} W_i^{(1)\prime} + \sum_{j=1}^{m} W_j^{(2)} W_j^{(2)\prime}\right]. \tag{13}$$

Choice of $g(.)$. If W is finitely supported, one can simply let $g(W) = W$. Yet, if W takes many values, it is convenient to group some of these values together, so that none of the $(\widehat{p}_k)_{k=1,...,K}$ is too small and the asymptotic framework below remains a good approximation. When W is not finitely supported, recall from Theorem 2 that $\overline{b}_d^g$ is sharp if $\nu_Y \perp\!\!\!\perp W|g(W)$ and $\nu_d \perp\!\!\!\perp W|g(W)$. Hence, we can expect tight bounds if $g(W)$ captures most of the dependence between (ν_Y, ν_d) and W. Since ν_Y and ν_d are already residuals, we seek to capture possible heteroskedasticity by regressing $|\nu_Y|$ and $|\nu_d|$ linearly on W. This yields two indices, $W'\widehat{\varsigma}_Y$ and $W'\widehat{\varsigma}_d$. The underlying idea is that if Y and η_d satisfy a linear location-scale model, namely $Y = W'\delta_Y + (W'\varsigma_Y)\xi_Y$ with $\xi_Y \perp\!\!\!\perp W$ and similarly for η_d, then $\nu_Y \perp\!\!\!\perp W|g(W)$ and $\nu_d \perp\!\!\!\perp W|g(W)$ hold with $g(W) = (W'\varsigma_Y, W'\varsigma_d)$. However, this construction does not ensure that g is finitely supported. To address this, we perform K-means clustering on $(W'\widehat{\varsigma}_Y, W'\widehat{\varsigma}_d)$. This yields a function g taking K values only. The choice of K is discussed in Section 5 below.

4.2 Asymptotic normality of $\widehat{\overline{b}}_d$ and inference on b_d

We now turn to the asymptotic properties of $\widehat{\overline{b}}_d$, and the construction of confidence intervals on b_d. For conciseness, we focus on the case without common variables; we briefly discuss the effect of these variables at the end of the section.

4.2.1 Asymptotic normality

We first establish the asymptotic normality of $\widehat{\overline{b}}_d$, under the following assumptions.

Assumption 2 *We observe $(Y_1, ..., Y_n)$ and $(X_{o,1}, ..., X_{o,m})$, two independent samples of i.i.d. variables with the same distribution as Y and X_o, respectively.*

Assumption 3 *One of the following holds:*

(i) $E[|Y|^{2+\varepsilon}] < \infty$ for some $\varepsilon > 0$, $|Supp(\eta_d)| = |Supp(X)| < \infty$ and $\forall h \in Supp(\eta_d)$, F^{-1} is continuous at $G(h)$.

(ii) $E[\|X\|^4] < \infty$, $|Supp(Y)| < \infty$, $Supp(\eta_d)$ is an interval and G is continuous.

(iii) $E(|Y|^{4+\varepsilon} + \|X\|^{4+\varepsilon}) < \infty$ *for some* $\varepsilon > 0$, F^{-1} *and* G *are continuous and for either* $Z = Y$ *or* $Z = \eta_d$, *the distribution of* Z *is continuous with respect to the Lebesgue measure and there exists* $C_1, C_2 > 0$ *such that for all* z *in the interior of Supp*(Z),

$$\frac{f_Z(z)}{F_Z(z)(1 - F_Z(z))} \geq C_1 \wedge \frac{C_2}{|z| \ln(1 + |z|)^2}. \tag{14}$$

We consider in Assumption 3 three possibilities, depending on whether η_d and Y are finitely supported or not. The first case corresponds to η_d being finitely supported. In such a case, Y can be continuous or discrete, as long as, in the latter case, there is no (h, y) such that $F(y) = G(h) \in (0, 1)$. The second case corresponds to Y being finitely supported and η_d continuous. The third case corresponds to the two variables being, loosely speaking, continuous (actually, case (iii) is compatible with Y having point masses, if we let $Z = \eta_d$). Then, we impose not only moment conditions but also (14). This condition holds on $\text{Supp}(Z) \cap [0, \infty)$ for all distributions that have increasing hazard rates, such as log-concave distributions (as their survival function is then log-concave). It also holds for many distributions with decreasing hazard rates, such as Pareto and Weibull distributions. More generally, we expect Condition (14) to be mild, since for any continuous probability measure μ with cdf F, density f and supremum of support equal to $\overline{x} \leq \infty$, we have, for all $A < \overline{x}$ satisfying $F(A) > 0$,

$$\int_A^{\overline{x}} \frac{f(x)}{F(x)(1 - F(x))} dx \geq \int_A^{\overline{x}} (-\ln[1 - F(x)])' dx = \infty.$$

On the other hand, for any $C_1, C_2 > 0$,

$$\int_A^{\overline{x}} C_1 \wedge \frac{C_2}{|x| \ln(1 + |x|)^2} dx < \infty.$$

Thus, one cannot have $f(x)/[F(x)(1 - F(x))] \leq C_1 \wedge C_2/(|x| \ln(1 + |x|)^2)$ for all x large enough; and similarly one cannot have $f(x)/[F(x)(1 - F(x))] \leq C_1 \wedge C_2/(|x| \ln(1 + |x|)^2)$ for all x small enough.

To define the asymptotic distribution, we introduce additional objects. First, let $h(x) := \int_0^1 F^{-1}[G(x^-) + u(G(x) - G(x^-))]du$ and

$$\begin{aligned}
\psi_1 &:= -\overline{b}_d(\eta_d^2 - E[\eta_d^2]), \\
\psi_2 &:= -E[h(\eta_d)T'_{-1}]E[T_{-1}T'_{-1}]^{-1}T_{-1}\eta_d, \\
\psi_3 &:= -\int [\mathbb{1}\{\eta_d \leq t\} - G(t)]F^{-1} \circ G(t)dt,
\end{aligned}$$

$$\psi_4 := -\int [\mathbb{1}\{Y \leq t\} - F(t)] G^{-1} \circ F(t) dt.$$

These four variables correspond to the influence functions of respectively $\sqrt{m}(\widehat{E}(\widehat{\eta}_d^2) - E(\eta_d^2))$, $\sqrt{m}\int_0^1 F^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt$, $\sqrt{m}\int_0^1 F^{-1}(G_m^{-1} - G^{-1})dt$, and $\sqrt{n}\int_0^1 G^{-1}(F_n^{-1} - F^{-1})dt$, with G_m the empirical cdf of the $(\eta_{dj})_{j=1,\ldots,m}$ (note that G_m cannot be computed in practice, since the $(\eta_{dj})_{j=1,\ldots,m}$ are unobserved).

Theorem 3 *Suppose that* $\min(m,n) \to \infty$, $n/(m+n) \to \lambda \in [0,1]$ *and Assumptions 1-3 hold. Then,*

$$\sqrt{\frac{mn}{m+n}}\left(\widehat{\overline{b}}_d - \overline{b}_d\right) \xrightarrow{d} \mathcal{N}(0, V_d),$$

where $V_d := \left[\lambda V(\psi_1 + \psi_2 + \psi_3) + (1-\lambda)V(\psi_4)\right] / E(\eta_d^2)^2$.

Remarks on the result. First, we comment on the assumptions underlying Theorem 3. We allow not only for $\lambda \in (0,1)$, but also for $\lambda = 0$ or $\lambda = 1$, which corresponds to cases where one sample is much larger than the other. In these cases, the asymptotic variance V_d simplifies. Also, when $\min(|\text{Supp}(X)|, |\text{Supp}(Y)|) < \infty$, we obtain weak convergence under minimal conditions; note that $E[\|X\|^4] < \infty$ is close to being necessary for the OLS estimator $\widehat{\gamma}$ of the regression of T_1 on T_{-1} to be $\sqrt{m}-$consistent.

When $\min(|\text{Supp}(X)|, |\text{Supp}(Y)|) = \infty$, the conditions we impose are probably not minimal, but note that a moment of order 4 for Y and η_d seems necessary in view of (12) and the discussion of Theorem 1 in Del Barrio et al. (2019). Moreover, closely related results in the literature on the asymptotic normality of $W_2(F_n, G_m)$ impose strong restrictions.[4] In particular, instead of Assumption 3-(iii), Proposition 2.3 in Del Barrio et al. (2019) imposes strong and high-level conditions (see (2-7)-(2.9) in their paper), while Theorem 14 in Berthet et al. (2020) also imposes strong regularity conditions. In particular, because their Assumption (FG) must hold for both the left and right tails of the distributions, one can show that their subconditions (FG1) and (FG3) already imply (up to letting $\varepsilon = 0$) Assumption 3-(iii) for both $Z = Y$ and $Z = \eta_d$.[5]

[4] By the proof of Point 2 of Theorem 3, we obtain, under Assumptions 1-3, the asymptotic normality of $(nm/(n+m))^{1/2}(W_2(F_n, G_m) - W_2(F, G))$.

[5] On the other hand, both Berthet et al. (2020) and Del Barrio et al. (2019) also consider more general Wasserstein distances than just W_2.

Sketch of the proof. In a first step, we account for the fact that η_d and $E[\eta_d^2]$ are estimated. This requires in particular to show that

$$\sqrt{m}\int_0^1 F_n^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt = -E[h(\eta_d)T'_{-1}]\sqrt{m}(\widehat{\gamma} - \gamma_0) + o_P(1),$$

where γ_0 is the limit in probability of $\widehat{\gamma}$. This result is not obvious; our proof relies in particular, again, on the Cambanis-Simons-Stout inequality. The second step is to study the asymptotic behavior of $(nm/(n+m))^{1/2}\int_0^1[F_n^{-1}(t)G_m^{-1}(t) - F^{-1}(t)G^{-1}(t)]dt$. Here, we use the decomposition

$$\begin{aligned}\int_0^1 F_n^{-1}(t)G_m^{-1}(t)dt = &\int_0^1 F^{-1}(t)(G_m^{-1}(t) - G^{-1}(t))dt \\ &+ \int_0^1 G^{-1}(t)(F_n^{-1}(t) - F^{-1}(t))dt + r_{n,m},\end{aligned}$$

where $r_{n,m} := \int_0^1(F_n^{-1}(t) - F^{-1}(t))(G_m^{-1}(t) - G^{-1}(t))dt$. We prove that the first two terms T_{1m} and T_{2n} are asymptotically linear by adapting results on L-statistics, see in particular Theorem 1 in Chapter 19 of Shorack and Wellner (1986). That the remainder term $r_{n,m}$ is negligible if Y (say) is finitely supported follows from the continuity of G^{-1} at the support points of Y. Note that if this continuity condition does not hold, we lose asymptotic normality; see Del Barrio et al. (2024) for the exact distribution in such cases. If Assumption 3-(iii) holds, we relate instead the remainder term to bounds on the convergence rate of $W_2(F_n, F)$ and $W_2(G_m, G)$. However, existing results on such rates, and in particular Theorem 1 in Fournier and Guillin (2015), are not sufficient for our purpose. Here, we improve upon their bound, which holds under weak restrictions, by leveraging in particular Condition (14). We do this by linking $W_2(F_n, F)$ to the variance of order statistics, and relying on a lemma similar to Corollary 2.12 in Boucheron and Thomas (2015); see Lemma 3 in Online Appendix E.3.

4.2.2 Confidence intervals

We construct confidence intervals on b_d using the asymptotic normality of $\widehat{\overline{b}}_d$ and a plug-in estimator of V_d. Specifically, let $\widehat{h}(x) = \int_0^1 F_n^{-1}[\widehat{G}_m(x^-) + u(\widehat{G}_m(x) - \widehat{G}_m(x^-))]du$ and

$$\begin{aligned}\widehat{\psi}_{1i} &:= -\widehat{\overline{b}}_d\left(\widehat{\eta}_{di}^2 - \frac{1}{m}\sum_{j=1}^m \widehat{\eta}_{dj}^2\right), \\ \widehat{\psi}_{2i} &:= -\left(\frac{1}{m}\sum_{j=1}^m \widehat{h}(\widehat{\eta}_{dj})T'_{-1j}\right)\left(\frac{1}{m}\sum_{j=1}^m T_{-1j}T'_{-1j}\right)^{-1} T_{-i}\widehat{\eta}_{di},\end{aligned}$$

$$\widehat{\psi}_{3i} := -\int [\mathbb{1}\{\widehat{\eta}_{di} \leq t\} - \widehat{G}_m(t)] F_n^{-1} \circ \widehat{G}_m(t) dt,$$
$$\widehat{\psi}_{4i} := -\int [\mathbb{1}\{Y_i \leq t\} - F_n(t)] \widehat{G}_m^{-1} \circ F_n(t) dt.$$

Then, define

$$\widehat{V}_d := \frac{1}{\left(\frac{1}{m}\sum_{j=1}^m \widehat{\eta}_{dj}^2\right)^2} \times \left[\frac{n}{m(n+m)} \sum_{j=1}^m \left(\widehat{\psi}_{1j} + \widehat{\psi}_{2j} + \widehat{\psi}_{3j}\right)^2 + \frac{m}{n(n+m)} \sum_{i=1}^n \widehat{\psi}_{4i}^2\right].$$

Note that $\widehat{V}_d$ depends on d; in particular, $\widehat{V}_{-d}$ is the estimator of the asymptotic variance of $\overline{b}_{-d} = -\underline{b}_d$. We then consider the following confidence intervals on b_d with nominal level $1-\alpha$:

$$\mathrm{CI}_{1-\alpha} := \left[-\widehat{\overline{b}}_{-d} - z_{1-\alpha}\sqrt{\frac{n+m}{nm}\widehat{V}_{-d}},\ \widehat{\overline{b}}_d + z_{1-\alpha}\sqrt{\frac{n+m}{nm}\widehat{V}_d}\right],$$

where $z_{1-\alpha}$ is the quantile of order $1-\alpha$ of a standard normal distribution. We can replace the usual quantile $z_{1-\alpha/2}$ by $z_{1-\alpha}$ since by Theorem 1, the identified interval of b_d is not reduced to a singleton ($\overline{b}_d > 0 > \underline{b}_d$) as long as $V(Y) > 0$.

Theorem 4 *Suppose that* $\min(m,n) \to \infty$, $n/(m+n) \to \lambda \in [0,1]$, *Assumptions 1-3 hold and* $V(Y) > 0$. *Then,*

$$\inf_{b_d \in [\underline{b}_d, \overline{b}_d]} \limsup_{n\to\infty} P\left(b_d \in CI_{1-\alpha}\right) = 1-\alpha.$$

Once again, the proof of Theorem 4 is not straightforward. In particular, two difficulties are (i) to prove convergence of $(1/m)\sum_{j=1}^m \widehat{h}(\widehat{\eta}_{dj}) T'_{-1j}$; (ii) to handle the terms including $\widehat{\psi}_{3i}$ and $\widehat{\psi}_{4i}$. For (ii), we rely in particular on an extension of Lemma A.1 in Del Barrio et al. (2019), see Lemma 4 in Online Appendix E.3.

4.2.3 Common variables

Given our focus on a finitely supported $g(W)$, the analysis is very similar to the case without common variables, so we mostly highlight the differences here, without providing a formal result for the sake of conciseness. First, the asymptotic variance of $\widehat{\overline{b}}_d$ includes additional terms due in particular to (i) the estimation of $\delta_d' E[WW']\delta_Y$; (ii) the estimation of the residual ν_Y. The exact expression of the asymptotic variance, which includes eleven terms instead of four as above, is given in Online Appendix B.

Then, the construction of the confidence interval is similar to that described above, with one important difference, which is to allow for the possibility of point identification. To maintain size control, we rely on Stoye (2020) to construct the confidence intervals. This method has the appealing features of not requiring any tuning parameter, being simple to compute, and relying on mild conditions, beyond the joint asymptotic normality of the lower and upper bounds. We implement this inference method in our Monte Carlo simulations (Section 5) and in the applications (Section 6).

5 Simulations

We now study the finite sample performances of our estimators and inference method. We consider a single DGP encompassing three cases of available data: one in which only Y and X_o are available, one in which X_c is also observed jointly and enters the main regression and one in which W_a, in addition to X_c, is observed. In the latter case, the parameters remain the same as in the second case. The DGP is as follows. We let $W_a \sim \mathcal{U}[0,1]$, $X_c \sim \mathcal{N}(0,1)$ and

$$\begin{aligned} X_o =& X_c a_1 + W_a a_2 + (1 + W_a d_1)\eta,\ \eta | W_a, X_c \sim \mathcal{N}(0, \sigma_\eta^2), \\ Y =& X_o b_1 + X_c b_2 + W_a d_2 + \varepsilon,\ \varepsilon | X, W_a, \eta \sim \mathcal{N}(0, \sigma_\varepsilon^2). \end{aligned}$$

We fix $a_1 = 1$, $a_2 = 10$, $d_1 = 1$, $\sigma_\eta = 1$, $b_1 = 1$, $b_2 = 1$, $d_2 = 0.25$ and $\sigma_\varepsilon = 4$. The true bounds in the first two cases are obtained by simulations, whereas there is a closed-form expression in the last case. We fix $n = m$ and vary it from 400 to 4,800. We construct $g(W)$ as described in Subsection 4.1.2, with $K = \max(2, \lfloor \min(n,m)^{0.2} \rfloor)$, where $\lfloor x \rfloor$ denotes the integer part of x; we discuss alternative choices of K below. The results are displayed in Table 1. We report the average of the estimated bounds ("Bounds") and the average of the estimated 95% confidence intervals $\text{CI}_{1-\alpha}$ ("95% CI") for $b^{0,1}$. We also report the mean difference between the length of the confidence sets and that of the identified set, see column "Ex. length" in the table. Finally, the column "Covg" corresponds to the minimum, over b_1 in the identified set of $b^{0,1}$, of the estimated probability that b_1 belongs to the confidence interval.

	Bounds	95% CI	Ex. length	Covg.
Panel 1:	Without (X_c, W_a)			
Identified	[-1.624,1.626]			
400	[-1.623,1.625]	[-1.743,1.746]	0.239	0.940
800	[-1.622,1.624]	[-1.707,1.709]	0.166	0.934
1,200	[-1.623,1.625]	[-1.693,1.695]	0.138	0.933
2,400	[-1.625,1.626]	[-1.674,1.676]	0.100	0.949
4,800	[-1.625,1.627]	[-1.66,1.662]	0.072	0.942
Panel 2:	With X_c			
Identified	[-1.583,1.585]			
400	[-1.563,1.566]	[-1.723,1.708]	0.264	0.941
800	[-1.573,1.575]	[-1.687,1.677]	0.196	0.956
1,200	[-1.574,1.576]	[-1.67,1.661]	0.163	0.954
2,400	[-1.578,1.581]	[-1.647,1.641]	0.120	0.959
4,800	[-1.579,1.582]	[-1.629,1.625]	0.086	0.966
Panel 3:	With (X_c, W_a)			
Identified	[0.196,1.405]			
400	[0.203,1.394]	[0.037,1.6]	0.354	0.963
800	[0.203,1.404]	[0.087,1.555]	0.259	0.955
1,200	[0.201,1.402]	[0.107,1.528]	0.211	0.952
2,400	[0.199,1.404]	[0.133,1.495]	0.153	0.953
4,800	[0.198,1.403]	[0.151,1.47]	0.109	0.964

Notes: results obtained with 2,000 simulations for each row. 400, 800 etc. correspond to the sizes of the two samples ($n = m$). Column "Bounds" reports either the identified set or the average of the estimated bounds over simulations. Column "95% CI" reports the average 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 "Covg." displays the minimum, over $b = (b_1, ..., b_p) \in \mathcal{B}$, of the estimated probability that $b_1 \in \mathrm{CI}_{1-\alpha}(b^{0,1})$. In Panel 1, the true coefficients of $(X_o, 1)$ are $(1.103, -0.393)$, while in Panels 2-3, the true coefficients of $(X_o, X_c, 1)$ are $(1.019, 0.981, 0.026)$.

Table 1: Monte Carlo simulation results on the confidence intervals for $b^{0,1}$

A couple of remarks are in order. First, as expected, the 95% confidence intervals shrink with the sample sizes n, approximately at the $n^{-1/2}$ rate in the three cases

we consider. This is reflected in the evolution of the excess length across sample sizes. Second, the confidence intervals exhibit satisfactory coverage. In particular, coverages for all panels are generally close to the nominal 95% level, even for small sample sizes. Coverage rates are generally conservative, but still very close to the nominal level for the specification reported in Panel 3. This is remarkable: one would in principle need to use the continuous variable $g(W) = 1 + W_a d_1$ to obtain the sharp bounds, by Theorem 2, whereas we instead rely on a finitely supported variable $g(W)$ with few points of support (from 3 to 5 when n varies from 400 to $4,800$). Third, and importantly, the identified set is much tighter in Panel 3 than in Panels 1 and 2. This illustrates the substantial identifying power of the auxiliary variable W_a. For this particular DGP, the identifying power - measured by the reduction in the length of the identified set - of W_a is in fact larger than that of the common regressor X_c.

Table 2 reports the computational time needed to compute the estimated bounds and associated confidence intervals. When W_a is observed, this time also includes the K-means clustering we perform to compute $g(W)$. The main takeaway is that our procedure is very fast: it takes less than 1 second when observing (X_c, W_a) with $n = m = 12,000$, and less than 12 seconds with $n = m$ as large as 120,000.

$n(= m)$	Without (X_c, W_a)	With X_c	With (X_c, W_a)
1,200	0.004	0.101	0.108
12,000	0.02	0.85	0.89
120,000	0.45	10.91	11.76

Notes: these times are obtained on the same DGP as above, taking the average over 100 replications and using our companion R package `RegCombinBLP`. We parallelize the computation over 20 CPUs on an Intel Xeon Gold 6130 CPU 2.10GHz with 382Gb of RAM.

Table 2: Time (in s.) for computing the point estimates and confidence intervals.

Finally, we explore the effect of the tuning parameter K on coverage; see Table 3 below, where we consider two sample sizes ($n = 1,200$ and $n = 6,000$). As expected, increasing K decreases the length of the CIs, but also reduces coverage. This probably reflects the fact that the estimated bounds become biased for larger K. On the other hand, coverage remains above 95% for $K = \max(2, \lfloor \min(n, m)^c \rfloor)$, $c < 1/3$, suggesting that our baseline choice of K with

$c = 0.2$ works well in practice.

Number of points K	Bounds	95% CI	Ex. length	Covg.
Panel 1:	$n = 1,200$			
Identified	[0.196,1.405]			
3	[0.200,1.405]	[0.105,1.531]	0.217	0.955
5	[0.203,1.401]	[0.109,1.526]	0.208	0.965
10	[0.210,1.394]	[0.114,1.515]	0.192	0.949
15	[0.217,1.388]	[0.120,1.508]	0.178	0.927
20	[0.221,1.381]	[0.124,1.499]	0.166	0.910
50	[0.257,1.344]	[0.158,1.459]	0.092	0.763
80	[0.277,1.327]	[0.174,1.444]	0.060	0.644
100	[0.275,1.330]	[0.172,1.447]	0.066	0.656
Panel 2:	$n = 6,000$			
Identified	[0.196,1.405]			
3	[0.196,1.406]	[0.155,1.466]	0.102	0.953
5	[0.196,1.404]	[0.155,1.463]	0.099	0.954
10	[0.200,1.404]	[0.159,1.463]	0.095	0.944
15	[0.200,1.401]	[0.157,1.458]	0.092	0.952
20	[0.202,1.400]	[0.160,1.457]	0.088	0.950
50	[0.209,1.390]	[0.166,1.445]	0.070	0.899
80	[0.217,1.384]	[0.174,1.438]	0.055	0.843
100	[0.222,1.379]	[0.178,1.433]	0.045	0.785

Notes: same DGP as above, observing (X_c, W_a). 3, 5, 10 etc. are the number of points K taken by $g(\cdot)$. Column "Bounds" reports either the identified set or the average of the estimated bounds over simulations. Column "95% CI" reports the average 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 "Covg." displays the minimum, over $b = (b_1, ..., b_p) \in \mathcal{B}$, of the estimated probability that $b_1 \in \text{CI}_{1-\alpha}(b^{0,1})$. The results are obtained with 1,000 simulations for each sample size.

Table 3: Simulation results when varying the number of points K taken by g.

6 Applications

We now illustrate our approach with two applications. We first study the influence of race on the probability of patent approval in the United States, revisiting recent work on this question (Dossi, 2024). We then investigate the relationship between

students' risk and time preferences and educational achievement across countries (Hanushek et al., 2022).

6.1 Race and patent approval

In our first application, we investigate the existence and magnitude of racial and ethnicity gaps in science and innovation. This question has attracted much interest in the recent empirical literature (see, e.g. Kerr, 2008; Antman et al., 2024; Dossi, 2024). A key challenge is that datasets typically do not measure race and ethnicity together with the outcome of interest. Using our notation, race/ethnicity is an outside regressor (X_o), with successful patent application being the outcome of interest (Y). Instead of X_o, we may observe other characteristics, such as the applicant's name. Then, in other datasets, we may observe these characteristics together with race and ethnicity. A commonly used strategy in this context is to impute race and ethnicity using applicant characteristics observed in both datasets. We take a different route and derive bounds that use both datasets without relying on the exclusion restriction implicit in the imputation approach.

Following Dossi (2024), we rely on two datasets. The first is the publicly available dataset released by the United States Patent and Trademark Office (USPTO) covering the universe of patent applications submitted in the United States. We use the Patent Examination research dataset (PatEx), which contains detailed information on all patent applications, including the full names of the applicants (Graham et al., 2015). We restrict the sample to applications filed between January 2001 and December 2018 and focus on utility patents.[6] We further restrict the sample to applicants based in the United States, and as in Dossi (2024), consider only the first inventor listed on the application.

We combine the PatEx dataset with data from the US Census. Namely, we use the information on the aggregate frequency of last names by race and ethnicity from the 2010 Decennial Census Surname Table (Comenetz, 2016). 6.3 million different last names were recorded for 295 million people.[7] Among them, we use the publicly released frequency by race and ethnicity of the 162,254 last names that occur more than 100 times, representing 90.1% of the overall population. We

[6]Utility patents, also referred to as "patents for invention", constitute 90% of the patent documents issued by the USPTO in recent years.

[7]In the following, we neglect the statistical uncertainty related to this sample.

consider in our analysis five different categories of race and ethnicity, namely: (i) Black or African American (11.99%), (ii) Asian and native Hawaiian and other Pacific Islander (4.86%), (iii) Hispanic or Latino (16.29%), (iv) American Indian or Alaska Native (1.76%), and (v) others, which includes White (64.40%) and those declaring to belong to two or more races (0.69%), and is used as our reference category.[8]

Estimation results are reported in Table 4, where the first column presents the TSTSLS point estimates, the second the point estimates of our bounds, and the last column reports the 95% confidence intervals computed from our asymptotic normality results. We use applicant's last name as an auxiliary variable (W_a). Our bounds correspond to $\overline{b}_d^g$ where, to reduce the size of the vector W, we define $g(W)$ as W_a unless the names W_a appear $L = 5$ times or less in the dataset of inventors, in which case we set $g(W) = 0$ (Table 7 in Appendix C.1 show that our results are robust to choosing $L = 3$ or $L = 10$ instead). Since inventors are a subset of the whole population, our bounds are plug-in estimates of Equation (11) above. In contrast to our bounds, the TSTSLS estimates rely on an exclusion restriction. Namely, the applicant's last names is assumed not to be predictive of patent approval once conditioning on applicant's race and ethnicity. While this type of name-based exclusion restriction has frequently been used in applied work, its validity is far from obvious in this particular context. In fact, it does not seem unreasonable to think that, in contrast to the TSTSLS exclusion restriction, any racial discrimination in the patent approval process would operate largely through the applicant's last name. This is consistent with the information available to patent examiners, who always observe applicants' last names but do not observe race, and only infrequently interview them in person (see Cockburn et al., 2002, and Avivi, 2024 for discussions of the USPTO selection process).[9]

Turning to the results, a key takeaway is that the TSTSLS results that are obtained using last names as an exclusion restriction are fragile. Notably, while the TSTSLS estimates point to Black inventors being significantly less likely to be granted a

[8]Estimation results are robust to splitting the "two or more races" category evenly across the other racial categories.

[9]One may argue that the TSTSLS identifies instead the effect of, e.g., having a Black- or Asian-sounding name. It is unclear whether this interpretation is warranted either. First, names could also predict other relevant characteristics. Second, this interpretation would require another exclusion restriction, namely that the coefficients of race in the "long" regression are zero, which is arguably strong as well.

patent, the bounds obtained with our method for this coefficient are wide, with a lower bound as large as -0.729 and an upper bound that is positive and large as well (0.360). While a similar conclusion holds for Hispanics and American natives, our bounds are somewhat more informative for the coefficient on Asians, with a lower bound of -0.137 and an upper bound of 0.187. At any rate, these results indicate that the conclusions one would reach from the TSTSLS estimates of significant racial differences in the probability of being granted a patent crucially hinge on the underlying exclusion restriction.

A final point is that, although the bounds reported in Table 4 tend to be wide, using last names as W_a does yield substantial improvements over the simple bounds based solely on Y and X_o. In particular, without W_a, the sharp lower bound for each of the four coefficients equals -1 and is therefore not informative.[10] Hence, even without exclusion restrictions, observing last names in both datasets delivers meaningful informational gains.

Coefficient	TSTSLS	Sharp bounds	95% CI
Black	-0.038	[-0.729, 0.360]	[-0.772, 0.383]
	(0.007)		
Hispanic	-0.032	[-0.559, 0.279]	[-0.572, 0.287]
	(0.005)		
Asian	0.041	[-0.137, 0.187]	[-0.145, 0.195]
	(0.002)		
American native	-0.047	[-0.801, 0.364]	[-0.831, 0.378]
	(0.005)		

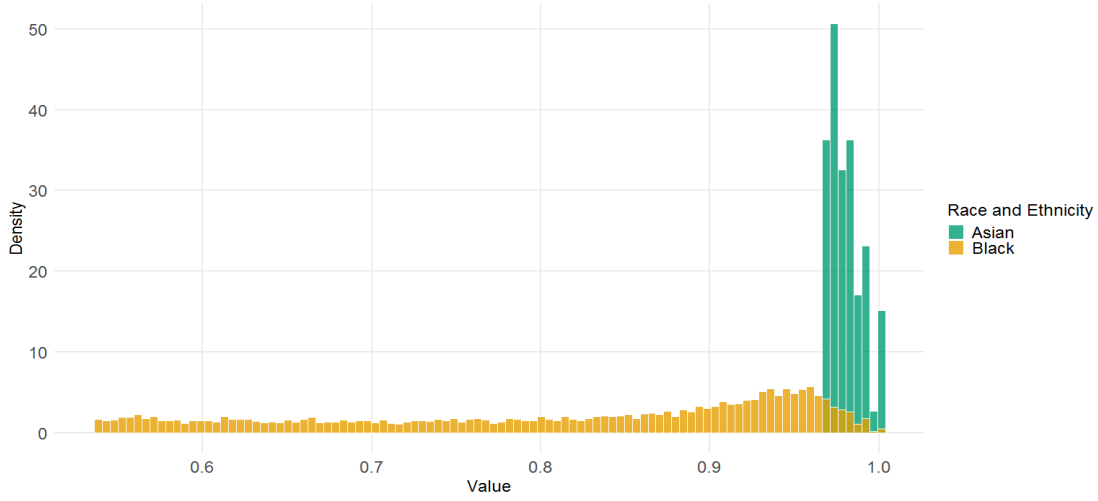
Notes: W_a = last names and no X_c, 2,146,799 patent applications and 91,055 names. The CIs and standard errors are obtained clustering at the last name level, following Dossi (2024).

Table 4: Estimation results for racial inequalities in patent approval

Next, we investigate why our bounds are more informative for some races/ethnicities than others. To do so, we report in Figure 1 the distribution of the racial frequen-

[10]This occurs here because (i) $P(Y = 0)$ is larger than P(race) for all other races than White and multiracial applicants, and ii) $P(Y = 1)$ is larger than P(White). The corresponding upper sharp bound is equal to 0.432 for each of the four coefficients. Again, this is due to the particular configurations of $P(Y = 1)$ and P(race). In our setup, 0.432 simply corresponds to $P(Y = 0)/P$(White).

cies conditional on last name, focusing on the names that are the most predictive of race and that together account for 10% of each sub-populations (here, Blacks and Asians - groups for which the bounds are relatively wide and more informative, respectively).This figure shows that last names are highly predictive of being Asian, much more so than for Blacks. This illustrates the connection between the informativeness of our bounds and the extent to which W_a (inventor's last name) is predictive of X_o (race/ethnicity).



Notes: 10% of the associated population after sorting by predictability represents 6,669 names for Blacks (dark orange), and 637 for Asians (dark green). We use 100 bins.

Figure 1: Distribution of racial frequencies for a given name for the most predictive names, which cumulatively account for 10% of the associated population

We conclude this analysis by exploring further the effect on our bounds of using more auxiliary information, as measured by the auxiliary variables W_a. Since the Census Surname table only contains racial characteristics associated with last names at the aggregate level, we cannot use this data for this purpose. Instead, we leverage the fact that voter registration data in North Carolina (Historical Voter Registration Snapshots) records historical individual data about active and inactive voters registered in North Carolina, with information about their full names, city, race and ethnicity. Thus, restricting to the set of inventors residing in North Carolina (62,112 applications associated with 23,689 unique inventors), we are able to merge application data with individual data from the Historical Voter

Registration Snapshots.[11]

Table 5 provides a comparison of different point estimates for our bounds, using different sets of W_a. A couple of comments are in order. The first one relates to the sample restrictions imposed by the common support requirements when using more comprehensive sets of W_a. The underlying reduction in sample size increases from around 3% when using last names only, to as much as 37% when using the complete name and city. Second, we only obtain very informative bounds in the latter case, in which we uniquely identify as much as 98% of the inventors. In other words, (very) high predictive power is needed to obtain tight bounds on the coefficients of interest.

W_a	Last name	Complete name	First, Last name, city	Complete name, city
Black	[-0.593, 0.327]	[-0.245, 0.122]	[-0.154, -0.004]	[-0.068, -0.043]
Hispanic	[-0.491, 0.329]	[-0.086, 0.115]	[-0.050, 0.079]	[0.021, 0.049]
Asian	[-0.269, 0.300]	[-0.075, 0.087]	[-0.056, 0.052]	[-0.013, 0.009]
American natives	[-0.676, 0.338]	[-0.270, 0.183]	[-0.253, -0.074]	[-0.106, -0.075]
Number applications	60,688	47,088	48,121	39,427
Number of inventors	22,904	16,164	16,710	12,418
Share of matched applic.	0.02	0.47	0.79	0.98

Notes: Complete name means first, last, and middle names. Total number of applications is 62,112 with 23,689 unique inventors in North Carolina. No X_c. The estimates are computed taking into account the share of matched applications, i.e. as a weighted average of the OLS coefficients obtained on the merged dataset of uniquely identified individuals based on the information W_a and the bounds obtained using the two datasets with unmatched observations.

Table 5: Bounds using different sets of W_a.

6.2 Preferences and educational achievement

Preference parameters, especially patience and risk taking, play an important role in human capital investment decisions. However, to our knowledge, no single data set jointly measures these preferences and test scores across countries. In the following, we build on the cross-country analysis of Hanushek et al. (2022) and combine data from the OECD's Programme for International Student Assessment

[11]To be representative of the population in North Carolina over the period 2001-2018, we actually use snapshots of 2006, 2013, and 2019, keeping only information about full names, city, and recorded race and ethnicity of all the uniquely identified active and inactive voters over this period. This data is openly available at www.ncsbe.gov/results-data/voter-history-data.

(PISA) with the Global Preference Survey (GPS) to examine how students' time and risk preferences are associated with educational achievement.

PISA assesses achievement in mathematics, science and reading for random samples of 15-year-old students on a three-year cycle, providing repeated cross-sectional data representative of each country-by-wave cell. In the following, we consider as our main "Y dataset" the standardized math test scores over the seven waves of PISA testing, covering the period 2000-2018. Over this period, a total of 86 countries participated at least once. We combine these test scores with data from the Global Preference Survey (see, e.g., Falk et al., 2018). The GPS provides scientifically validated data on several preference parameters from representative samples, of around 1,000 respondents in each country surveyed in 2012, measuring patience, risk taking, positive and negative reciprocity, altruism, and trust (X_o), for 49 different countries. The GPS also records gender for each respondent, which we use as a common regressor X_c, together with the country. Restricting the analysis to this subset of 49 countries yields test score data for a total of 1,992,276 students.[12]

In Table 6, we compare our bounds on the coefficients of patience and risk taking with the TSTSLS estimates considered by Hanushek et al. (2022), where both variables are imputed using country dummies. We consider alternative specifications depending on whether only patience, only risk taking or both are included in the regression. In Panel D, we also include other preference variables (positive and negative reciprocity, altruism and trust) as controls in the regression. Hence, in this last specification, X_o is of dimension 6.

A couple of comments are in order. A first takeaway is that, in contrast to the previous application and despite the absence of any W_a here, our bounds tend to be informative. This holds for both the coefficients of patience and risk-taking, and across all four specifications reported in the table. That the bounds remain informative is particularly noteworthy in Panel D, where we control for four additional preference parameters all included in X_o. One might indeed have expected that increasing the dimension of X_o would cause the bounds to widen substantially, yet this is not the case here.

[12]See Appendix C.2 for more details on the GPS dataset, especially on the measurements of preference parameters. As the PISA and GPS datasets are representative of the same common population after reweighting the observations by the corresponding survey weights, we use the survey weights in our analysis.

	Panel A			Panel B		
	TSTSLS	Sharp bnd.	CI on $b^{0,1}$	TSTSLS	Sharp bnd.	CI on $b^{0,1}$
Patience	0.898	[-0.844,0.943]	[-1.035,1.187]	-	-	-
	(0.045)			-	-	-
Risk-taking	-	-	-	-0.596	[-1.018,0.858]	[-1.188,1.021]
	-	-	-	(0.128)		
	Panel C			**Panel D**		
	TSTSLS	Sharp bnd.	CI on $b^{0,1}$	TSTSLS	Sharp bnd.	CI on $b^{0,1}$
Patience	1.172	[-0.842,0.986]	[-1.105,1.223]	1.122	[-0.853,0.977]	[-1.083,1.236]
	(0.046)			(0.050)		
Risk-taking	-1.311	[-1.067,0.877]	[-1.259,1.051]	-1.345	[-1.094,0.918]	[-1.301,1.036]
	(0.108)			(0.128)		
Additional controls				X	X	X
Tests equality	Stat.	p.value		Stat.	p.value	
U. bnd. Patience	1.114	0.132		0.801	0.211	
L. bnd. Risk-taking	1.530	0.063		1.419	0.077	

Notes: X_c includes countries and gender dummies, no W_a here. Dependent variable: PISA math test score in all PISA waves 2000–2018. Respectively 1,992,276 and 49,689 observations for the PISA and GPS datasets. Least squares regression weighted by students' sampling probability. Additional preference controls are positive and negative reciprocity, altruism and trust. The standard errors of the TSTSLS estimators, our confidence intervals and the tests of equality between our lower or upper bounds and the TSTSLS estimators take into account the clustering at the country level.

Table 6: Preferences and Student Math Achievement across Countries

Related to this, for Panels C and D, the TSTSLS estimates of the coefficients associated with patience and risk-taking both lie outside of the estimated sharp bounds, and, for the risk-taking coefficient, outside of the 95% confidence intervals as well. One-sided tests of equality between the lower bound of our identified set on the coefficient of risk-taking and the TSTSLS point estimate in Panels C and D leads us to reject this hypothesis at the 10% level, consistent with a violation of the underlying exclusion restrictions. Recall that the TSTSLS estimator relies on the arguably strong assumption that countries do not affect test scores beyond their effects through risk aversion and patience. In terms of magnitudes, focusing on Panel D where we include additional preference controls, our bounds indicate that a 1 standard deviation (SD) increase in patience is at most associated with 0.977 SD increase in math test scores, against 1.122 SD using TSTSLS. Similarly, it follows from our bounds that a 1 SD increase in risk-taking is, at most, associated with a decline of 1.094 SD in student achievement, against a larger decline of

-1.345 using TSTSLS.

Finally and importantly for practice, our inference method can be implemented at a low computational cost. Even though the datasets used in this analysis contain a very large number of observations (1,992,276 and 49,689), it takes 5 minutes only to reproduce the results of Panels A and B, 9 minutes for Panel C, and 21.5 minutes for Panel D (where X_o is of dimension 6), using our R package `RegCombinBLP`.[13]

7 Conclusion

We study regression coefficients in a context where the outcome of interest and some of the covariates are observed in two different datasets that cannot be matched. This type of data combination environment arises very frequently in various empirical setups. The usual approach, which consists in imputing the outcome Y or the outside regressors X_o using auxiliary variables W_a, hinges on exclusion restrictions that may not hold in practice. We take a different route and derive sharp bounds on the regression coefficients using only the observed distributions. As they take a simple form, these bounds can be estimated at a low computational cost; we also derive simple and easy-to-compute confidence intervals.

We illustrate our method with two applications. The first studies racial disparities in patent approval, the second the effects of patience and risk-taking on test scores. The first application highlights that in some cases, results based on an imputation approach crucially rely on the underlying exclusion restriction; without it, uncertainty on the true coefficients of interet remains large. The second application shows that our bounds can be informative on the magnitude of the effects, and can also lead to reject the imputation-based approach.

[13]We parallelize the computation over 15 CPUs on an Intel Xeon Gold 6130 CPU 2.10GHz with 382Gb of RAM.

References

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

Antman, F. M., K. B. Doran, X. Qian, and B. A. Weinberg (2024). Demographic diversity and economic research: Fields of specialization and research on race, ethnicity, and inequality. National Bureau of Economic Research working paper.

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

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2024). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. National Bureau of Economic Research working paper.

Avivi, H. (2024). Are patents examiners gender neutral? Mimeo.

Berthet, P., J.-C. Fort, and T. Klein (2020). A central limit theorem for wasserstein type distances between two distinct univariate distributions. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 56*(2), 954 – 982.

Bobkov, S. and M. Ledoux (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, Volume 261 (1259). American Mathematical Society.

Bogachev, V. I. (2007). *Measure theory*. Springer.

Bontemps, C., J.-P. Florens, and N. Meddahi (2025). Functional ecological inference. *Journal of Econometrics 248*, 105918.

Boucheron, S. and M. Thomas (2015). Tail index estimation, concentration and adaptivity. *Electronic Journal of Statistics 9*(2), 2751–2792.

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

Cambanis, S., G. Simons, and W. Stout (1976). Inequalities for $\mathscr{E}k(X, Y)$ when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 36*(4), 285–294.

Cockburn, I. M., S. Kortum, and S. Stern (2002). Are all patent examiners equal? the impact of characteristics on patent statistics and litigation outcomes. National Bureau of Economic Research working paper.

Comenetz, J. (2016). Frequently occurring surnames in the 2010 census. *United States Census Bureau*, 1–8.

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

Crossley, T. F., P. Levell, and S. Poupakis (2022). Regression with an imputed dependent variable. *Journal of Applied Econometrics 37*(7), 1277–1294.

Del Barrio, E., A. González Sanz, and J.-M. Loubes (2024). Central limit theorems for semi-discrete wasserstein distances. *Bernoulli 30*(1), 554–580.

Del Barrio, E., P. Gordaliza, and J.-M. Loubes (2019). A central limit theorem for lp transportation cost on the real line with application to fairness assessment in machine learning. *Information and Inference: A Journal of the IMA 8*(4), 817–849.

Delon, J., N. Gozlan, and A. Saint Dizier (2023). Generalized wasserstein barycenters between probability measures living on different subspaces. *The Annals of Applied Probability 33*(6A), 4395–4423.

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

Dossi, G. (2024). Race and science. Working Paper.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2025). Partially linear models under data combination. *Review of Economic Studies 92*(1), 238–267.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics 133*(4), 1645–1692.

Falk, A., A. Becker, T. Dohmen, D. Huffman, and U. Sunde (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science 69*(4), 1935–1950.

Falkner, N. and G. Teschl (2012). On the substitution rule for lebesgue–stieltjes integrals. *Expositiones Mathematicae 30*(4), 412–418.

Fan, Y., H. Park, B. Pass, and X. Shi (2025). Partial identification in moment models with incomplete data-a conditional optimal transport approach. arXiv preprint arXiv:2503.16098v2.

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

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

Fournier, N. and A. Guillin (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields 162*(3), 707–738.

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

Graham, S. J., A. C. Marco, and R. Miller (2015). The USPTO patent examination research dataset: A window on the process of patent examination. Georgia Tech Scheller College of Business Research Paper No. WP.43.

Hanushek, E. A., L. Kinne, P. Lergetporer, and L. Woessmann (2022). Patience, risk-taking, and human capital investment across countries. *Economic Journal 132*(646), 2290–2307.

Hwang, Y. (2025). Bounding omitted variable bias using auxiliary data: With an application to estimate neighborhood effects. Working Paper.

Jakubowski, A. (2021). A complement to the Chebyshev integral inequality. *Statistics & Probability Letters 168*, 108934.

Kerr, W. (2008). Ethnic scientific communities and international technology diffusion. *Review of Economics and Statistics 90*(3), 518–537.

Kitawaga, T. and M. Sawada (2023). Linear regressions, shorts to long. Institute of Economic Research Hitotsubashi University, Discussion Paper Series A No.747.

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

Martin, G. J. and A. Yurukoglu (2017). Bias in cable news: Persuasion and polarization. *American Economic Review 107*(9), 2565–2599.

Meango, R., M. Henry, and I. Mourifie (2025). Combining stated and revealed preferences. arXiv preprint arXiv:2507.13552.

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

Moon, S. (2024). Partial identification of individual-level parameters using aggregate data in a nonparametric binary outcome model. arXiv preprint arXiv:2403.07236.

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

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

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

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine 8*(4), 431–440.

Rambachan, A., R. Singh, and D. Viviano (2024). Program evaluation with remotely sensed outcomes. *arXiv preprint arXiv:2411.10959*.

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

Rubner, Y., C. Tomasi, and L. J. Guibas (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision 40*, 99–121.

Santavirta, T. and J. Stuhler (2024). Name-based estimators of intergenerational mobility. *The Economic Journal 134*(663), 2982–3016.

Shorack, G. and J. Wellner (1986). *Empirical Processes with Applications to Statistics*. SIAM, Classics in Applied Mathematics.

Stoye, J. (2020). A simple, short, but never-empty confidence interval for partially identified parameters. arXiv preprint arXiv:2010.10484.

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

van der Vaart, A. W. and J. A. Wellner (2023). *Weak convergence and empirical processes*. Springer.

Villani, C. (2009). *Optimal transport: old and new*, Volume 338. Springer.

A Comparison with Pacini (2019)

A.1 Sharpness

Pacini (2019) gives the expression of $\overline{b}_d$, also allowing for common regressors (denoted by z in his paper) but not for additional variables W_a. His bounds coincide with ours when X_o is univariate, but not otherwise. In the multidimensional case, his expression of $\overline{b}_d$ is an upper bound of the true bound. This is so because the equality in Lemma 5 of Pacini (2019) should be replaced by an inequality.

To see this, first remark that $\mathcal{F}$ there is the set of cdfs $(F_{1y}, ..., F_{d_x y})$ that are compatible with the distributions of (x, z) and (y, z), with F_{ky} denoting the joint cdf of (x_k, y). Hence, in the third equality "$F_{ky} \in \mathcal{F}$" is not well-defined. A natural fix is then to replace it by "$F_{ky} \in \mathcal{F}_k$", where $\mathcal{F}_k$ denotes the set of cdfs F_{ky} compatible with the laws of (x, z) and (y, z). But then, the third equality in the proof of Lemma 5 does not hold, because $\mathcal{F}$ is not a cartesian product of $\mathcal{F}_k$ in general: it is instead a (strict in general) subset of the cartesian product.

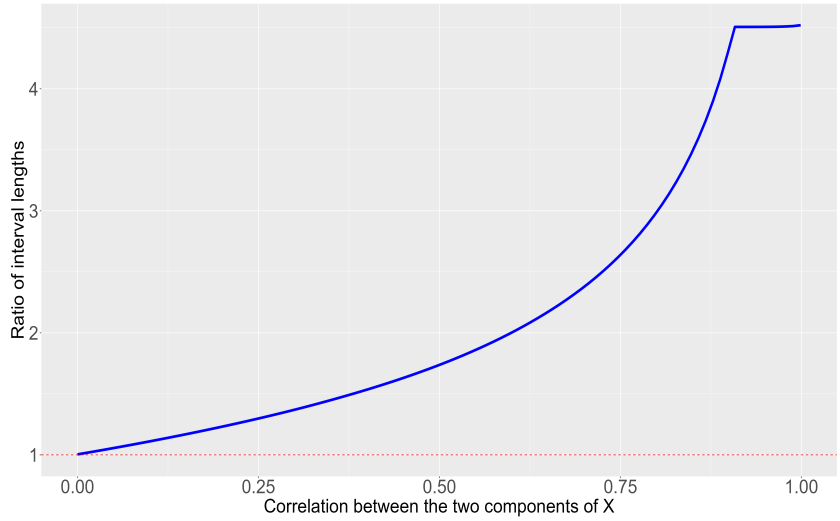
A.2 Numerical comparison

We illustrate in the following that the bounds provided in Pacini (2019) can in practice be substantially larger than the sharp bounds. To this end, we consider

the following class of DGPs, indexed by ρ: $\log(Y) \sim \mathcal{N}(0, 2)$ and $X = (X_1, X_2) \sim \mathcal{N}(0, \Sigma)$ with Σ defined as in (8) (note that Σ depends on ρ). To compare the two types of bounds, we consider the following ratio

$$R := \frac{\overline{b}_d^p - \underline{b}_d^p}{\overline{b}_d - \underline{b}_d},$$

where $d = (1, 0)'$ and $(\underline{b}_d^p, \overline{b}_d^p)$ denote Pacini's bounds. Figure 2 reports R as a function of ρ. When X_1 and X_2 are independent, the two intervals coincide, but the sharp bounds become tighter as the correlation between X_1 and X_2 increases. With $\rho \geq 0.88$, the sharp identification interval is more than four times shorter than the one obtained with Pacini's bounds.



Notes: results obtained by approximating the true bounds using a sample of size 10^5. The ratio of interval lengths is the ratio of the intervals obtained using Pacini (2019) bounds and the sharp bounds.

Figure 2: Comparison between Pacini (2019) bounds and the sharp bounds

B Asymptotic variance with common variables

With common variables, the asymptotic variance takes the form

$$\begin{aligned} V_d^g :=& \frac{\lambda}{E[\eta_d^2]} V(\psi_1^g + \psi_{2,2}^g + \psi_4^g + \psi_5^g + \psi_8^g + \psi_{10}^g) \\ &+ \frac{(1-\lambda)}{E[\eta_d^2]} V(\psi_{2,1}^g + \psi_3^g + \psi_6^g + \psi_7^g + \psi_9^g), \end{aligned}$$

where $\psi_1^g := \delta_Y' \left[W^{(2)}\nu_d - E[W^{(2)}T_{-1}']E[T_{-1}T_{-1}']^{-1}T_{-1}\eta_d\right]$, $\psi_{2,1}^g := \delta_d' \left(W^{(1)}W^{(1)\prime} - E[W^{(1)}W^{(1)\prime}]\right)\delta_y$, $\psi_{2,2}^g := \delta_d' \left(W^{(2)}W^{(2)\prime} - E[W^{(2)}W^{(2)\prime}]\right)\delta_y$, $\psi_3^g := \delta_d' W^{(1)}\nu_Y$, $\psi_{10}^g := -\bar{b}_d(\eta_d^2 - E[\eta_d^2])$ and

$$\begin{aligned}
\psi_4^g &:= -\left(\sum_{k=1}^{K} \mathbb{1}\left\{g(W^{(2)}) = g_k\right\} E\left[F_{|k}^{-1} \circ G_{|k}(\nu_d)W^{(2)\prime} | g(W^{(2)}) = g_k\right]\right) E[W^{(2)}W^{(2)\prime}]^{-1} \\
&\quad \times \left[W^{(2)}\nu_d - E[W^{(2)}T_{-1}']E[T_{-1}T_{-1}']^{-1}T_{-1}\eta_d\right], \\
\psi_5^g &:= -\sum_{k=1}^{K} \mathbb{1}\left\{g(W^{(2)}) = g_k\right\} \int [\mathbb{1}\{\nu_d \leq t\} - G_{|k}(t)]F_{|k}^{-1} \circ G_{|k}(t)dt, \\
\psi_6^g &:= -\left(\sum_{k=1}^{K} \mathbb{1}\left\{g(W^{(1)}) = g_k\right\} E\left[G_{|k}^{-1} \circ F_{|k}(\nu_Y)W^{(1)\prime} | g(W^{(1)}) = g_k\right]\right) \\
&\quad \times E[W^{(1)}W^{(1)\prime}]^{-1}W^{(1)}\nu_Y, \\
\psi_7^g &:= -\sum_{k=1}^{K} \mathbb{1}\left\{g(W^{(1)}) = g_k\right\} \int [\mathbb{1}\{\nu_Y \leq t\} - F_{|k}(t)]G_{|k}^{-1} \circ F_{|k}(t)dt, \\
\psi_8^g &:= \sum_{k=1}^{K} (\mathbb{1}\left\{g(W^{(2)}) = g_k\right\} - p_k) \int_0^1 F_{|k}^{-1}(u)G_{|k}^{-1}(u)du, \\
\psi_9^g &:= \sum_{k=1}^{K} (\mathbb{1}\left\{g(W^{(1)}) = g_k\right\} - p_k) \int_0^1 F_{|k}^{-1}(u)G_{|k}^{-1}(u)du.
\end{aligned}$$

C Additional elements on the applications

C.1 Additional results on the first application

	Limit $L = 3$		Limit $L = 5$ (baseline)		Limit $L = 10$	
Coefficient	Sharp bounds	95% CI	Sharp bounds	95% CI	Sharp bounds	95% CI
Black	[-0.718, 0.357]	[-0.762, 0.379]	[-0.729, 0.360]	[-0.772, 0.383]	[-0.751, 0.368]	[-0.793, 0.390]
Hispanic	[-0.529, 0.269]	[-0.541, 0.276]	[-0.559, 0.279]	[-0.572, 0.287]	[-0.615, 0.299]	[-0.623, 0.301]
Asian	[-0.129, 0.184]	[-0.138, 0.192]	[-0.137, 0.187]	[-0.145, 0.195]	[-0.154, 0.194]	[-0.163, 0.202]
American native	[-0.791, 0.361]	[-0.822, 0.375]	[-0.801, 0.364]	[-0.831, 0.378]	[-0.821, 0.371]	[-0.851, 0.385]

Notes: W_a = last names and no X_c, 2,146,799 patent applications and 91,055 names. The CIs and standard errors are obtained clustering at the last name level, following Dossi (2024). We define $g(W)$ as W_a unless the name W_a appears L times or less in the dataset of inventors, in which case $g(W) = 0$. We report results for $L = 3$, $L = 5$ (baseline), and $L = 10$.

Table 7: Robustness of estimation results for the racial inequalities on access to patents according to the definition of $g(W)$.

C.2 Additional details on the second application

The GPS survey (see Falk et al., 2018) covers the following countries: Argentina, Australia, Austria, Bosnia and Herzegovina, Brazil, Canada, Switzerland, Chile, Colombia, Costa Rica, Czech Republic, Germany, Algeria, Spain, Estonia, Finland, France, United Kingdom, Georgia, Greece, Croatia, Hungary, Indonesia, Israel, Italy, Jordan, Japan, Kazakhstan, South Korea, Lithuania, Morocco, Moldova, Mexico, Netherlands, Peru, Philippines, Poland, Portugal, Romania, Russia, Saudi Arabia, Serbia, Sweden, Thailand, Turkey, Ukraine, United States, Vietnam, United Arab Emirates. The preference measures therein are based on 12 survey items, which are summarized in Table I in Falk et al. (2018). We gather them below with their respective weights for completeness:

1. Patience: Intertemporal choice sequence using staircase method (0.712); Self-assessment: willingness to wait (0.288).
2. Risk taking: Lottery choice sequence using staircase method (0.473); Self-assessment: willingness to take risks in general (0.527).
3. Positive reciprocity: Gift in exchange for help (0.515); Self-assessment: willingness to return a favor (0.485).
4. Negative reciprocity: Self-assessment: willingness to take revenge (0.374); Self-assessment: willingness to punish unfair behavior toward self (0.313); Self-assessment: willingness to punish unfair behavior toward others (0.313).
5. Altruism: Donation decision (0.635); Self-assessment: willingness to give to good causes (0.365).
6. Trust: Self-assessment: people have only the best intentions (1).

The weights endogenously emerged from a preliminary experimental validation procedure (see Falk et al., 2023). Each preference measure is standardized at the individual level.

D Proofs of the identification results

D.1 Theorem 1

First, if $b \in \mathcal{B}$, then there exists r.v. $(\widetilde{Y}, \widetilde{X})$ with $F_{\widetilde{X}} = F_X$, $F_{\widetilde{Y}} = F_Y$ and such that $EL(\widetilde{Y}|\widetilde{X}) = \widetilde{X}'b$. Thus, $\widetilde{\varepsilon} := \widetilde{Y} - \widetilde{X}'b$ satisfies $E(\widetilde{\varepsilon}) = 0$ and $\text{Cov}(\widetilde{X}, \widetilde{\varepsilon}) = 0$. Hence, $E[\widetilde{Y}] = E[\widetilde{X}'b]$ and

$$V(\widetilde{Y}) = V(\widetilde{X}'b) + V(\widetilde{\varepsilon}) \geq V(\widetilde{X}'b).$$

As a result, $E(Y) = E(X'b)$, $V(Y) \geq V(X'b)$ and $\mathcal{B} \subseteq \mathcal{E}$. This also implies that $\mathcal{B}$ is bounded.

Now, let us prove that $\mathcal{B}$ is closed. This, in turn, will imply that $\mathcal{B}$ is compact. Let $b_n \in \mathcal{B}$ for all $n \geq 1$ with $b_n \to b$ and let us prove that $b \in \mathcal{B}$. Let $(\widetilde{X}_n, \widetilde{Y}_n)$ such that $F_{\widetilde{X}_n} = F_X$, $F_{\widetilde{Y}_n} = F_Y$ and $b_n = E(\widetilde{X}_n \widetilde{X}_n')^{-1} E(\widetilde{X}_n \widetilde{Y}_n)$. Since $E(\widetilde{X}_n \widetilde{X}_n') = E(XX')$, it suffices to prove that there exists $(\widetilde{X}, \widetilde{Y})$, with $F_{\widetilde{X}} = F_X$, $F_{\widetilde{Y}} = F_Y$, such that $E[\widetilde{X}\widetilde{Y}] = c := E(XX')b$. First, note that for all M,

$$\begin{aligned} P\left(\|(\widetilde{X}_n, \widetilde{Y}_n)\| \geq M\right) &\leq \frac{E\left[\|(\widetilde{X}_n, \widetilde{Y}_n)\|^2\right]}{M^2} \\ &\leq \frac{E[\|X\|^2] + E[Y^2]}{M^2}. \end{aligned}$$

Hence, $(\widetilde{X}_n, \widetilde{Y}_n)$ is uniformly tight. Then, by Prokhorov's theorem, there exists a subsequence $(\widetilde{X}_{n_j}, \widetilde{Y}_{n_j})$ that converges in distribution, to $(\widetilde{X}, \widetilde{Y})$ say. Moreover, $F_{\widetilde{X}} = F_X$ and $F_{\widetilde{Y}} = F_Y$. Now, remark that for all $(x, y) \in \mathbb{R}^{+2}$ and all $M > 0$, we have

$$xy\mathbb{1}\{xy > M\} \leq x^2\mathbb{1}\left\{x > M^{1/2}\right\} + y^2\mathbb{1}\left\{y > M^{1/2}\right\}.$$

As a result, for all $n \geq 1$ and all $M > 0$,

$$\begin{aligned} &E\left[\|\widetilde{X}_{n_j}\widetilde{Y}_{n_j}\|\mathbb{1}\left\{\|\widetilde{X}_{n_j}\widetilde{Y}_{n_j}\| > M\right\}\right] \\ \leq &E\left[\|\widetilde{X}_{n_j}\|^2\mathbb{1}\left\{\|\widetilde{X}_{n_j}\| > M^{1/2}\right\}\right] + E\left[\widetilde{Y}_{n_j}^2\mathbb{1}\left\{|\widetilde{Y}_{n_j}| > M^{1/2}\right\}\right] \\ = &E\left[\|X\|^2\mathbb{1}\left\{\|X\| > M^{1/2}\right\}\right] + E\left[Y^2\mathbb{1}\left\{|Y| > M^{1/2}\right\}\right]. \end{aligned}$$

As a result, by the dominated convergence theorem, $\widetilde{X}_{n_j}\widetilde{Y}_{n_j}$ is asymptotically uniformly integrable. This implies (see, e.g. van der Vaart, 2000, Theorem 2.20) that

$$E\left[\widetilde{X}_{n_j}\widetilde{Y}_{n_j}\right] \to E[\widetilde{X}\widetilde{Y}].$$

Because we also have $E\left[\widetilde{X}_{n_j}\widetilde{Y}_{n_j}\right] \to c$, we finally obtain $E[\widetilde{X}\widetilde{Y}] = c$. This proves that $\mathcal{B}$ is closed.

Next, we prove that $\mathcal{B}$ is convex. Let $(b_1, b_2) \in \mathcal{B}^2$ and fix $p \in [0, 1]$. Then, there exists $(\widetilde{X}_1, \widetilde{Y}_1)$ and $(\widetilde{X}_2, \widetilde{Y}_2)$ rationalizing respectively b_1 and b_2. Let D following a Bernoulli distribution with probability p, $D \sim$Be(p), independent of these random variables and let $(\widetilde{Y}, \widetilde{X}) = (\widetilde{Y}_1, \widetilde{X}_1)$ if $D = 1$, $(\widetilde{Y}, \widetilde{X}) = (\widetilde{Y}_2, \widetilde{X}_2)$ otherwise. Then, $F_{\widetilde{X}} = F_X$, $F_{\widetilde{Y}} = F_Y$ and

$$E\left[\widetilde{X}\widetilde{Y}\right] = pE\left[\widetilde{X}_1\widetilde{Y}_1\right] + (1-p)E\left[\widetilde{X}_2\widetilde{Y}_2\right]$$

$$=E(XX')(pb_1+(1-p)b_2).$$

Hence, $EL(\widetilde{Y}|\widetilde{X})=\widetilde{X}'(pb_1+(1-p)b_2)$, which implies that $\mathcal{B}$ is convex.

Now, we prove $\overline{b}_d=E[F^{-1}_{d'E[XX']^{-1}X}(U)F_Y^{-1}(U)]$. We have

$$\overline{b}_d=\max_{\Pi\in\mathcal{M}(F_X,F_Y)}\int\left[d'E[XX']^{-1}x\right]\,y\,d\Pi(x,y), \tag{15}$$

where $\mathcal{M}(F,G)$ denotes the set of probability measures with marginal cdfs equal to F and G. Remark that for any $c=(c_1,...,c_p)$ and any $(\widetilde{X},\widetilde{Y})\sim\Pi\in\mathcal{M}(F_X,F_Y)$,

$$(c\widetilde{X},\widetilde{Y})\sim\Pi\in\mathcal{M}(F_{cX},F_Y).$$

Therefore, letting $X_d:=d'E[XX']^{-1}X$, we obtain

$$\overline{b}_d\leq\max_{\Pi\in\mathcal{M}(F_{X_d},F_Y)}\int uyd\Pi(u,y).$$

Moreover, by the Cambanis-Simons-Stout inequality (see Cambanis et al., 1976),

$$\max_{\Pi\in\mathcal{M}(F_{X_d},F_Y)}\int uyd\Pi(u,y)=E[F^{-1}_{X_d}(U)F_Y^{-1}(U)]. \tag{16}$$

Hence, $\overline{b}_d\leq E[F^{-1}_{X_d}(U)F_Y^{-1}(U)]$.

Now, for any $U\sim\mathcal{U}([0,1])$, let $\widetilde{Y}=F_Y^{-1}(U)$. Let also C denote a copula of $M'E[XX']^{-1}X$ (recall the construction of M at the beginning of Section 3.1) and let $(U_2,...,U_p)$ be uniform random variables such that $(U,U_2,...,U_p)$ has cdf equal to C. Let us define

$$S_d=(F^{-1}_{X_d}(U),F^{-1}_{d_2'E[XX']^{-1}X}(U_2),...,F^{-1}_{d_p'E[XX']^{-1}X}(U_p))'.$$

By construction, $S_d\sim M'E[XX']^{-1}X$. Then, let $\widetilde{X}=(M'E[XX']^{-1})^{-1}S_d$, so that $\widetilde{X}\sim X$. Let Π^* denote the distribution of $(\widetilde{X},\widetilde{Y})$. We have $\Pi^*\in\mathcal{M}(F_X,F_Y)$. Moreover,

$$d'E[XX']^{-1}\widetilde{X}=d'M'^{-1}S_d=F^{-1}_{X_d}(U),$$

where the last equality follows since $e'_{1,p}\times M'=d'$. Thus, by definition of $\overline{b}_d$, $\overline{b}_d\geq E[F^{-1}_{X_d}(U)F_Y^{-1}(U)]$. Equation (3) follows.

Next, we prove (4). It suffices to show that $X_d=\eta_d/E(\eta_d^2)$. Remark that

$$d'E(XX')^{-1}X=e'_{1,p}M'E(XX')^{-1}M(M^{-1}X)=e'_{1,p}E(TT')^{-1}T.$$

Moreover, $\eta_d = \gamma' T$, with $\gamma := [1, -E(T_1 T_{-1})' E(T_{-1} T_{-1}')^{-1}]'$. Thus,

$$E(\eta_d^2) = \gamma' E(TT')\gamma = E(T_1^2) - E(T_1 T_{-1})' E(T_{-1} T_{-1}')^{-1} E(T_1 T_{-1}).$$

As a result, $E(TT') \times \gamma / E(\eta_d^2) = e_{1,p}$. The result follows since then,

$$X_d = e_{1,p}' E(TT')^{-1} T = \gamma' T / E(\eta_d^2) = \eta_d / E(\eta_d^2).$$

Finally, if $V(Y) > 0$, $F_Y^{-1}(U)$ is not constant. By Assumption 1, we also have $V(\eta_d) > 0$ and thus $F_{\eta_d}^{-1}(U)$ is not constant either. Then, by Theorem 1.1 and 1.2 of Jakubowski (2021) and using $F_{\eta_d}^{-1}(U) \sim \eta_d$ and $E[\eta_d] = 0$, we obtain

$$E[F_{\eta_d}^{-1}(U) F_Y^{-1}(U)] > E[F_{\eta_d}^{-1}(U)] E[F_Y^{-1}(U)] = 0.$$

The last point of the theorem follows.

D.2 Theorem 2

Since $b^0 = E(XX')^{-1} E(XY)$, the exact same reasoning as in the proof of Theorem 1 shows that the identified set $\mathcal{B}(w)$ of $E(XX')^{-1} E(XY|W = w)$ is convex. By integrating over w, $\mathcal{B}$ is thus convex. It is also bounded as a subset of $\mathcal{E}$.

Let's now prove that $\mathcal{B}_d = [\underline{b}_d, \overline{b}_d]$, with $\overline{b}_d$ satisfying (6); this will also imply that $\mathcal{B}$ is compact. By the same reasoning as in the proof of Theorem 1, we also have that the identified set $\mathcal{B}_d(w)$ of $b'd$ for $b \in \mathcal{B}(w)$ is the identified set of $E[\eta_d Y|W = w]/E(\eta_d^2)$ and that $\mathcal{B}_d(w) = [\underline{b}_d(w), \overline{b}_d(w)]$, with $\overline{b}_d(w) := \sup\{b'd : b \in \mathcal{B}(w)\}$. Let U be such that $U|W$ is uniformly distributed on $[0, 1]$. Then, $\overline{b}_d(w)$ satisfies

$$\begin{aligned}\overline{b}_d(w) =& \frac{1}{E(\eta_d^2)} E\left[F^{-1}_{W'\delta_d + \nu_d|W}(U|W) F^{-1}_{W'\delta_Y + \nu_Y|W}(U|W)|W = w\right] \\ =& \frac{1}{E(\eta_d^2)} E\left[\left(W'\delta_d + F^{-1}_{\nu_d|W}(U|W)\right)\left(W'\delta_Y + F^{-1}_{\nu_Y|W}(U|W)\right)|W = w\right].\end{aligned}$$

Next, we have $b_d = E[\eta_d Y]/E(\eta_d^2)$ by construction and $E[\eta_d Y] = E[E[\eta_d Y|W]] \leq E[\overline{b}_d(W)] E(\eta_d^2)$. Moreover, the bound is reached by considering

$$(\eta_d, Y) = (F^{-1}_{W'\delta_d + \nu_d|W}(U|W), F^{-1}_{W'\delta_Y + \nu_Y|W}(U|W)).$$

Thus, $\overline{b}_d = E[\overline{b}_d(W)]$. Since $(W, F^{-1}_{\nu_\ell|W}(U|W))$ (with $\ell \in \{d, Y\}$) has the same distribution as (W, ν_ℓ), we obtain

$$\overline{b}_d = \frac{1}{E(\eta_d^2)} \left\{\delta_d' E\left[WW'\right]\delta_Y + E\left[F^{-1}_{\nu_d|W}(U|W) W'\delta_Y\right] + E\left[F^{-1}_{\nu_Y|W}(U|W) W'\delta_d\right]\right\}$$

$$
\begin{aligned}
&+\frac{1}{E(\eta_d^2)}E\left[F^{-1}_{\nu_d|W}(U|W)F^{-1}_{\nu_Y|W}(U|W)\right]\\
=&\frac{1}{E(\eta_d^2)}\left\{\delta_d' E\left[WW'\right]\delta_Y + E\left[F^{-1}_{\nu_d|W}(U|W)F^{-1}_{\nu_Y|W}(U|W)\right]\right\}.
\end{aligned}
$$

To prove that $\bar{b}_d \leq \bar{b}_d^g$, remark that

$$
\begin{aligned}
E[\eta_d Y] =& E\left[(W'\delta_d + \nu_d)(W'\delta_Y + \nu_Y)\right]\\
=&\delta_d E\left[WW'\right]\delta_Y + E[\nu_d\nu_Y]\\
=&\delta_d E\left[WW'\right]\delta_Y + E\left[E[\nu_d\nu_Y|g(W)]\right]\\
\leq&\delta_d E\left[WW'\right]\delta_Y + E\left[F^{-1}_{\nu_d|g(W)}(U|g(W))F^{-1}_{\nu_Y|g(W)}(U|g(W))\right],
\end{aligned}
$$

where the last inequality follows by the Cambanis-Simons-Stout inequality. If $\nu_d \perp\!\!\!\perp W|g(W)$ and $\nu_Y \perp\!\!\!\perp W|g(W)$, the last expression is equal to $\bar{b}_d$. The third point of the proposition follows.

D.3 Proposition 1

Let us denote by $\mathcal{B}_Z(w)$ the identified set of $E[X_dY|W=w]$ when observing Z, whereas $\mathcal{B}(w)$ still denotes the identified set of $E[X_dY|W=w]$ without the knowledge of Z. Again, the same reasoning as in the proof of Theorem 1 shows that the identified set $\mathcal{B}_Z(w)$ of $E[X_dY|W=w]$ is non-empty, closed and convex. Thus, it is characterized by its so-called support function $\sigma_{\mathcal{B}_Z(w)}(d) := \sup\{b'd : b \in \mathcal{B}_Z(w)\}$. As in (15), we have

$$
\sigma_{\mathcal{B}_Z(w)}(d) = \max_{\Pi\in\mathcal{M}(F_{W,X_o},F_{W,Y,Z})}\int\left[d'E[XX']^{-1}(x_o', x_c')'\right]\, y\, d\Pi(w,x_o,y,z),
$$

where $w = (x_c', w_a')'$. By Lemma 3.3 of Delon et al. (2023),

$$
\begin{aligned}
\sigma_{\mathcal{B}_Z(w)}(d) =& \max_{\Pi\in\mathcal{M}(F_{W,X_o},F_{W,Y})}\int\left[d'E[XX']^{-1}(x_o', x_c')'\right]\, y\, d\Pi(w,x_o,y)\\
=&\sigma_{\mathcal{B}(w)}(d),
\end{aligned}
$$

the support function of $\mathcal{B}(w)$ evaluated at d. Hence, by integrating over w, we obtain $\sigma_{\mathcal{B}_Z} = \sigma_{\mathcal{B}}$. The result follows since these functions characterize $\mathcal{B}_Z$ and $\mathcal{B}$.

Online Appendix

E Proofs of the statistical results

E.1 Theorem 3

We prove the results in three main steps. The first step obtains linear approximations of terms related to the estimation of η_d and $E(\eta_d^2)$. The second step obtains a linear approximation of two other terms. The third step shows that a remainder term is negligible and concludes.

1. Linear approximation of the first terms

We first show that

$$\sqrt{\frac{nm}{n+m}}\left(\widehat{\overline{b}}_d - \overline{b}_d\right) = \frac{1}{E(\eta_d^2)}\left[\sqrt{\frac{nm}{n+m}}\int_0^1 (F_n^{-1}G_m^{-1} - F^{-1}G^{-1})dt + \frac{\sqrt{\lambda}}{m^{1/2}}\sum_{i=1}^m (\psi_{1i} + \psi_{2i})\right] + o_P(1). \tag{17}$$

First, remark that

$$\widehat{\overline{b}}_d - \overline{b}_d = \frac{1}{\widehat{E}(\widehat{\eta}_d^2)}\left[\int_0^1 (F_n^{-1}\widehat{G}_m^{-1} - F^{-1}G^{-1})dt - \overline{b}_d(\widehat{E}(\widehat{\eta}_d^2) - E(\eta_d^2))\right]. \tag{18}$$

Moreover, since $\widehat{\eta}_{di} - \eta_{di} = -T'_{-1i}(\widehat{\gamma} - \gamma_0)$,

$$\begin{aligned}\widehat{E}(\widehat{\eta}_d^2) - E(\eta_d^2) &= \frac{1}{m}\sum_{i=1}^m \eta_{di}^2 - E[\eta_d^2] - \frac{2}{m}\sum_{i=1}^m \eta_{di}T'_{-1i}(\widehat{\gamma} - \gamma_0) \\ &\quad + (\widehat{\gamma} - \gamma_0)'\left(\frac{1}{m}\sum_{i=1}^m T_{-1i}T'_{-1i}\right)(\widehat{\gamma} - \gamma_0) \\ &= \frac{1}{m}\sum_{i=1}^m \eta_{di}^2 - E[\eta_d^2] + o_P(m^{-1/2}).\end{aligned}$$

The last equality follows since $E[\|X\|^4] < \infty$ implies both $\widehat{\gamma} - \gamma_0 = O_P(m^{-1/2})$ and $(1/m)\sum_{i=1}^m \eta_{di}T_{-1i} \xrightarrow{P} 0$. Combined with (18), $n/(n+m) \to \lambda$ and the definition of ψ_1, this yields

$$\sqrt{\frac{nm}{n+m}}\left(\widehat{\overline{b}}_d - \overline{b}_d\right) = \frac{1}{E(\eta_d^2)}\left[\sqrt{\frac{nm}{n+m}}\int_0^1 (F_n^{-1}\widehat{G}_m^{-1} - F^{-1}G^{-1})dt + \frac{\sqrt{\lambda}}{m^{1/2}}\sum_{i=1}^m \psi_{1i}\right] + o_P(1). \tag{19}$$

Let us now prove that

$$\sqrt{m}\int_0^1 F_n^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt = -E[h(\eta_d)T'_{-1}]\sqrt{m}(\widehat{\gamma} - \gamma_0) + o_P(1). \qquad (20)$$

When combined with (19), the standard result that

$$\sqrt{m}(\widehat{\gamma} - \gamma_0) = E[T_{-1}T'_{-1}]^{-1}\frac{1}{m^{1/2}}\sum_{i=1}^{m} T_{-1i}\eta_{di} + o_P(1),$$

and the definition of ψ_2, this will entail (17).

Remark that if $\eta_d \sim G$ and $U|X \sim \mathcal{U}[0,1]$, then

$$G(\eta_d^-) + U(G(\eta_d) - G(\eta_d^-)) \sim \mathcal{U}[0,1].$$

As a result, if we let $(U_1, ..., U_m)$ be i.i.d., uniform variables, the $\tilde{Y}_i := F^{-1}[G(\eta_{di}^-) + U_i(G(\eta_{di}) - G(\eta_{di}^-))]$ are i.i.d. with cdf F. Let σ_1 denote a permutation on $\{1, ..., m\}$ such that $\eta_{\sigma_1(1)} \leq ... \leq \eta_{\sigma_1(m)}$ and $\tilde{Y}_{\sigma_1(1)} \leq ... \leq \tilde{Y}_{\sigma_1(m)}$. Let also σ_2 denote a permutation on $\{1, ..., m\}$ such that $\widehat{\eta}_{\sigma_2(1)} \leq ... \leq \widehat{\eta}_{\sigma_2(m)}$; if σ_1 satisfies these inequalities, let $\sigma_2 := \sigma_1$. Finally, let $\lceil \cdot \rceil$ denote the ceiling function. Then, define $Q_m(t) := \eta_{d\sigma_2(\lceil mt \rceil)}$ and $\widehat{Q}_m(t) := \widehat{\eta}_{d\sigma_1(\lceil mt \rceil)}$. By the Cambanis-Simons-Stout inequality,

$$\int_0^1 F_n^{-1}(\widehat{Q}_m - G_m^{-1})dt \leq \int_0^1 F_n^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt \leq \int_0^1 F_n^{-1}(\widehat{G}_m^{-1} - Q_m)dt.$$

Next, remark that

$$\begin{aligned}
\widehat{Q}_m(t) - G_m^{-1}(t) &= -T'_{-1\sigma_1(\lceil mt \rceil)}(\widehat{\gamma} - \gamma_0), \\
\widehat{G}_m^{-1}(t) - Q_m(t) &= -T'_{-1\sigma_2(\lceil mt \rceil)}(\widehat{\gamma} - \gamma_0).
\end{aligned}$$

Then, letting $Q_{1m}(t) := T_{-1\sigma_1(\lceil mt \rceil)}$ and $Q_{2m}(t) := T_{-1\sigma_2(\lceil mt \rceil)}$, we obtain

$$\begin{aligned}
-\left[\int_0^1 F_n^{-1}Q'_{1m}dt\right](\widehat{\gamma} - \gamma_0) &\leq \int_0^1 F_n^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt \\
&\leq -\left[\int_0^1 F_n^{-1}Q'_{2m}dt\right](\widehat{\gamma} - \gamma_0). \qquad (21)
\end{aligned}$$

Let $\tilde{F}_m$ denote the empirical cdf of $(\tilde{Y}_i)_{i=1,...,m}$. Then,

$$\begin{aligned}
\left(\int_0^1 \left[F_n^{-1} - \tilde{F}_m^{-1}\right]^2 dt\right)^{1/2} &= W_2(F_n, \tilde{F}_m) \\
&\leq W_2(F_n, F) + W_2(\tilde{F}_m, F) \\
&\xrightarrow{P} 0.
\end{aligned}$$

The inequality holds since W_2 is a distance. The convergence to 0 follows since convergence of the Wasserstein-2 distance is equivalent to weak convergence and convergence of the second moment (see, e.g., Theorem 6.9 in Villani, 2009), and both $(Y_i)_{i=1,...,n}$ and $(\tilde{Y}_i)_{i=1,...,m}$ are i.i.d. with cdf F. Hence, we have, for $k \in \{1, 2\}$

$$\begin{aligned}\left\| \int_0^1 \left(F_n^{-1} - \tilde{F}_m^{-1} \right) Q'_{km} dt \right\| &\leq \left(\int_0^1 \left[F_n^{-1} - \tilde{F}_m^{-1} \right]^2 dt \right)^{1/2} \left(\int_0^1 \|Q_{km}\|^2 \, dt \right)^{1/2} \\ &= o_P(1). \end{aligned} \tag{22}$$

Next, remark that

$$\begin{aligned}\int_0^1 \tilde{F}_m^{-1} Q_{1m} dt &= \frac{1}{m} \sum_{i=1}^m \tilde{Y}_{\sigma_1(i)} T_{-1\sigma_1(i)} \\ &= \frac{1}{m} \sum_{i=1}^m \tilde{Y}_i T_{-1i} \\ &\xrightarrow{P} E[\tilde{Y}_1 T_{-11}].\end{aligned}$$

Moreover, by definition of h and $\tilde{Y}$,

$$E[\tilde{Y} T_{-1}] = E[E[\tilde{Y}|\eta_d, T_{-1}]T_{-1}] = E[h(\eta_d)T_{-1}].$$

Together with (22), this proves that

$$\int_0^1 F_n^{-1} Q_{1m} dt \xrightarrow{P} E[h(\eta_d)T_{-1}].$$

Using (22) again but with $k = 2$ and (21), (20) follows provided that

$$\int_0^1 \tilde{F}_m^{-1} Q_{2m} dt \xrightarrow{P} E[h(\eta_d)T_{-1}]. \tag{23}$$

Let us now show that (23) hold under Assumption 3. First, consider the case where Assumption 3-(i) holds. Because $|\text{Supp}(\eta_d)| = |\text{Supp}(X)|$, we have, by (55) in Lemma 2 and with probability approaching one (wpao), $\widehat{\eta}_{di} > \widehat{\eta}_{dj}$ that implies $\eta_{di} > \eta_{dj}$ for all $i \neq j$. Thus, because $\eta_{d\sigma_1(i)} \leq \eta_{d\sigma_1(i+1)}$, we have $\widehat{\eta}_{d\sigma_1(i)} \leq \widehat{\eta}_{d\sigma_1(i+1)}$ for all $i = 1, ..., m-1$. By construction of σ_2, this implies that $\sigma_2 = \sigma_1$, wpao. Since $\int_0^1 \tilde{F}_m^{-1} Q_{2m} dt = \int_0^1 \tilde{F}_m^{-1} Q_{1m} dt$ in this case, we obtain (23).

Next, assume that Assumption 3-(ii) holds. Let $\text{Supp}(Y) = \{y_1, ..., y_K\}$ with $-\infty =: y_0 < y_1 < ... < y_K$. Then $\tilde{Y}_i = F^{-1}(G(\eta_{di}))$. Moreover, F^{-1} is constant on $I_k := (F(y_{k-1}), F(y_k)]$ for all $k \in \{1, ..., K\}$. By (55) and continuity of G, we have

$$P\left(\exists (i, k, \ell) \in \{1, ..., m\} \times \{1, ..., K\}^2 : \, k \neq \ell, \, G(\eta_{di}) \in I_k, G(\widehat{\eta}_{di}) \in I_\ell \right) \to 0.$$

As a result, $\tilde{Y}_i = F^{-1}(G(\widehat{\eta}_{di}))$ for all i wpao. Under this event, $\sigma_2 = \sigma_1$ and as above, we obtain (23).

Finally, suppose that Assumption 3-(iii) holds. Then, $h(x) = F^{-1}[G(x)]$ is continuous and $\tilde{Y}_i = h(\eta_{di})$ for all i. Let $\widehat{\tilde{Y}}_i := h(\widehat{\eta}_{di})$. For any $M > 0$, let $f_M(x) = \min(\max(x, -M), M)$. Fix $\delta > 0$. By the dominated convergence theorem

$$\lim_{K\to\infty} E[h(\eta_d)^2 \mathbb{1}\{|\eta_d| > K\}] \to 0.$$

Then, let $K > 0$ be such that

$$E[h(\eta_d)^2 \mathbb{1}\{|\eta_d| > K\}] < \frac{\delta}{6E[\|T_{-1}\|^2]}.$$

Let also $M = \max(h(K), -h(-K))$. Then,

$$\begin{aligned}
\int_0^1 \tilde{F}_m^{-1} Q_{2m} dt &= \frac{1}{m}\sum_{i=1}^m \tilde{Y}_{\sigma_1(i)} T_{-1\sigma_2(i)} \\
&= \frac{1}{m}\sum_{i=1}^m \left[\tilde{Y}_{\sigma_1(i)} - f_M(\tilde{Y}_{\sigma_1(i)})\right] T_{-1\sigma_2(i)} + \frac{1}{m}\sum_{i=1}^m f_M(\widehat{\tilde{Y}}_{\sigma_2(i)}) T_{-1\sigma_2(i)} \\
&+ \frac{1}{m}\sum_{i=1}^m [f_M(\tilde{Y}_{\sigma_1(i)}) - f_M(\widehat{\tilde{Y}}_{\sigma_2(i)})] T_{-1\sigma_2(i)} \\
&=: T_0 + T_1 + T_2.
\end{aligned}$$

Consider T_0. By Cauchy-Schwarz inequality,

$$\|T_0\| \leq \left(\frac{1}{m}\sum_{i=1}^m \left[\tilde{Y}_i - f_M(\tilde{Y}_i)\right]^2\right)^{1/2} \left(\frac{1}{m}\sum_{i=1}^m \|T_{-1i}\|^2\right)^{1/2}. \tag{24}$$

Since h is increasing, $|\eta_{di}| \leq K$ implies $|\tilde{Y}_i| \leq M$. Then, $\tilde{Y}_i - f_M(\tilde{Y}_i) \neq 0$ implies $|\eta_{di}| > K$. Remark also that $|x - f_M(x)| \leq |x|$. Then,

$$|\tilde{Y}_i - f_M(\tilde{Y}_i)| \leq |h(\eta_{di})| \mathbb{1}\{|\eta_{di}| > K\}.$$

As a result,

$$\frac{1}{m}\sum_{i=1}^m \left[\tilde{Y}_i - f_M(\tilde{Y}_i)\right]^2 \leq \frac{1}{m}\sum_{i=1}^m h(\eta_{di})^2 \mathbb{1}\{|\eta_{di}| > K\}.$$

By the law of large numbers (LLN) and definition of K, we obtain, with probability approaching one (wpao),

$$\|T_0\| \leq \frac{\delta}{5}. \tag{25}$$

Next, consider T_1. We have

$$\begin{aligned} T_1 &= \frac{1}{m}\sum_{i=1}^{m} f_M(\widehat{\tilde{Y}}_i)T_{-1i} \\ &= \frac{1}{m}\sum_{i=1}^{m}\left[f_M(\widehat{\tilde{Y}}_i) - f_M(\tilde{Y}_i)\right]T_{-1i} + \frac{1}{m}\sum_{i=1}^{m}\tilde{Y}_i T_{-1i} + \frac{1}{m}\sum_{i=1}^{m}[f_M(\tilde{Y}_i) - \tilde{Y}_i]T_{-1i} \\ &=: T_{11} + T_{12} + T_{13}. \end{aligned}$$

By the LLN, wpao,

$$\|T_{12} - E[h(\eta_{d1})T_{-11}]\| \leq \frac{\delta}{5}. \tag{26}$$

By Cauchy-Schwarz inequality, we obtain for T_{13} the same inequality as (24). Thus, wpao,

$$\|T_{13}\| \leq \frac{\delta}{5}. \tag{27}$$

Turning to T_{11}, we have, by Cauchy-Schwarz inequality,

$$\|T_{11}\| \leq \left\{\frac{1}{m}\sum_{i=1}^{m}\left[f_M(\widehat{\tilde{Y}}_i) - f_M(\tilde{Y}_i)\right]^2\right\}^{1/2}\left\{\frac{1}{m}\sum_{i=1}^{m}\|T_{-1i}\|^2\right\}^{1/2}. \tag{28}$$

Remark that $\min(|\widehat{\eta}_{di}|, |\eta_{di}|) \geq K$ implies $\min(|\widehat{\tilde{Y}}_i|, |\tilde{Y}_i|) \geq M$, and then $f_M(\widehat{\tilde{Y}}_i) = f_M(\tilde{Y}_i)$. Then,

$$\begin{aligned} \left|f_M(\widehat{\tilde{Y}}_i) - f_M(\tilde{Y}_i)\right| &\leq \left|\widehat{\tilde{Y}}_i - \tilde{Y}_i\right| \mathbb{1}\left\{\min(|\widehat{\eta}_{di}|, |\eta_{di}|) < K\right\} \\ &= |h(\widehat{\eta}_{di}) - h(\eta_{di})|\, \mathbb{1}\left\{\min(|\widehat{\eta}_{di}|, |\eta_{di}|) < K\right\}. \end{aligned}$$

Because $I := [-K, K]$ is compact, there exists $\nu > 0$ such that for all $(x, y) \in I^2$, $|x - y| < \nu$ implies $|h(x) - h(y)| < \delta/\{6E[\|T_{-1}\|^2]\}$. Given (55) in Lemma 2, (28) and the LLN, we have, wpao

$$\|T_{11}\| \leq \frac{\delta}{5}. \tag{29}$$

Finally, consider T_2. First, by Cauchy-Schwarz inequality,

$$T_2 \leq \left\{\frac{1}{m}\sum_{i=1}^{m}\left[f_M(\tilde{Y}_{\sigma_1(i)}) - f_M(\widehat{\tilde{Y}}_{\sigma_2(i)})\right]^2\right\}^{1/2}\left\{\frac{1}{m}\sum_{i=1}^{m}\|T_{-1i}\|^2\right\}^{1/2}. \tag{30}$$

By the rearrangement inequality, because $f_M \circ h$ is increasing,

$$\sum_{i=1}^{m} f_M(\tilde{Y}_{\sigma_1(i)}) f_M(\widehat{\tilde{Y}}_{\sigma_2(i)}) \geq \sum_{i=1}^{m} f_M(\tilde{Y}_i) f_M(\widehat{\tilde{Y}}_i).$$

Thus, by what precedes, we have, wpao,

$$\frac{1}{m}\sum_{i=1}^{m}\left[f_M(\tilde{Y}_{\sigma_1(i)}) - f_M(\widehat{\tilde{Y}}_{\sigma_2(i)})\right]^2 \leq \frac{1}{m}\sum_{i=1}^{m}\left[f_M(\tilde{Y}_i) - f_M(\widehat{\tilde{Y}}_i)\right]^2$$

$$\leq \frac{\delta^2}{6E[\|T_{-1}\|^2]}.$$

When combined with (30) and the LLN, we have, wpao

$$\|T_2\| \leq \frac{\delta}{5}. \tag{31}$$

Finally, by combining (25)-(28) and (31), we obtain that wpao,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \tilde{Y}_{\sigma_1(i)} T_{-1\sigma_2(i)} - E[h(\eta_d) T_{-1}] \right\| \leq \delta.$$

Because δ was arbitrary, (23), and in turn (20), follows.

2. Linear approximation of the other terms

Consider the following decomposition

$$\int_0^1 F_n^{-1} G_m^{-1} dt = \int_0^1 F^{-1}(G_m^{-1} - G^{-1}) dt + \int_0^1 G^{-1}(F_n^{-1} - F^{-1}) dt + r_{n,m},$$

where $r_{n,m} := \int_0^1 (F_n^{-1} - F^{-1})(G_m^{-1} - G^{-1}) dt$. We prove that the first two terms T_{1m} and T_{2n} are asymptotically linear. We prove below that the last term is asymptotically negligible.

First, consider $T_{2n} = \int_0^1 G^{-1}(F_n^{-1} - F^{-1}) dt$. We can always construct i.i.d. uniform random variables ξ_i such that $Y_i = F^{-1}(\xi_i)$, see e.g. Eq. (55) p.57 in Shorack and Wellner (1986, SW hereafter). Now, we apply Theorem 19.1 in SW, combined with their Remark 2 p.667. Remark that their $\tilde{T}_n$ defined in their Eq. (56) corresponds to our $\int_0^1 G^{-1} F_n^{-1} dt$, with their h being the identity function so that their $g(\mathbb{G}_n^{-1})$ is our F_n^{-1} and their J is our G^{-1}. Given that their (58) is the same as their (11), with just $\Psi_n = \Psi$, we can replace in their Theorem 19.1-(i), provided that their Assumptions 19.1 and 19.2 hold, $T_n - \mu_n$ by their $\tilde{T}_n - \mu$, which is our T_{2n}.

We first check that SW's Assumption 19.1 holds. First, assume that Assumption 3-(i) holds. Then, by Remark 19.1 in SW, we have $|F^{-1}(t)| \leq M_1/[t(1-t)]^{1/(2+\varepsilon)}$ for some M_1. Moreover, $|G^{-1}(t)| \leq \max(\text{Supp}(\eta_d))$. Thus, (16) and (19) in SW's Assumption 19.1 holds, with their (b_1, b_2, d_1, d_2) satisfying $b_1 = b_2 = 0$ and $d_1 = d_2 = 1/(2+\varepsilon)$ and thus their a satisfying $a < 1/2$. If instead Assumption 3-(ii) holds, we reason similarly but instead use $b_1 = b_2 = 1/4$ and $d_1 = d_2 = 0$. Finally, assume that Assumption 3-(iii) holds. Using again Remark 19.1 in SW, (16) and (19) in their Assumption 19.1 holds, with their (b_1, b_2, d_1, d_2) satisfying $b_1 = ... = d_2 = 1/(4+\varepsilon)$ and thus their a satisfying again $a < 1/2$.

Now, let us check that SW's Assumption 19.2 holds. Since $J_n = J$ in $\tilde{T}_n$, their assumption reduces in our context to the continuity of G^{-1} except on a set of μ-measure 0, where μ is the measure associated with F^{-1}. If Assumption 3-(i) holds, G^{-1} is continuous except at $G(h)$ for $h \in \text{Supp}(\eta_d)$. But since F^{-1} is continuous at $G(h)$, $\mu(\{G(h)\}) = 0$. If, instead, Assumption 3-(ii) holds, the fact that $\text{Supp}(\eta_d)$ is an interval implies that G^{-1} is continuous. Finally, if Assumption 3-(iii) holds, because G^{-1} is monotone, its set of discontinuities $\mathcal{D}_{G^{-1}}$ is countable. Since F^{-1} is continuous, $\mu(\{x\}) = 0$ for each $x \in \mathcal{D}_{G^{-1}}$. Hence, in all cases, SW's Assumption 19.2 holds.

Then, by Theorem 19.1 in SW and their equation just above (13),

$$\sqrt{n}T_{2n} = -\frac{1}{n^{1/2}}\sum_{i=1}^{n}\int_0^1 [\mathbb{1}\{\xi_i \leq t\} - t]G^{-1}(t)dF^{-1}(t) + o_P(1).$$

Using $m/(n+m) \to (1-\lambda)$, Lemma 1 below and the definition of ψ_4, we obtain

$$\sqrt{\frac{nm}{n+m}}T_{2n} = (1-\lambda)^{1/2}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\psi_{4i} + o_P(1). \tag{32}$$

We reason similarly for T_{1m}. Note that Assumption 3 is not symmetric in F and G so care should be taken when checking Assumption 19.2 in SW (their Assumption 19.1 holds by just reverting the choices of (b_1, b_2) and (d_1, d_2) considered above). Now, we must check the continuity of F^{-1} except on a set of μ-measure 0, where μ is the measure associated with G^{-1}. If Assumption 3-(i) holds, note that the set of discontinuity points $\mathcal{D}_{F^{-1}}$ of F^{-1} is countable. Moreover, by assumption, $\mathcal{D}_{F^{-1}} \cap G(\text{Supp}(\eta_d)) = \emptyset$. Hence, $\mu(\{x\}) = 0$ for each $x \in \mathcal{D}_{F^{-1}}$, implying that $\mu(\mathcal{D}_{F^{-1}}) = 0$. If Assumption 3-(ii) holds, the same holds because G^{-1} is continuous. Finally, if Assumption 3-(iii) holds, F^{-1} is continuous so Assumption 19.2 in SW directly holds. Hence, at the end of the day, we obtain

$$\sqrt{\frac{nm}{n+m}}T_{1m} = \lambda^{1/2}\frac{1}{m^{1/2}}\sum_{i=1}^{m}\psi_{3i} + o_P(1). \tag{33}$$

3. Asymptotically negligible remainder term

We now show that $R_{n,m} := \sqrt{nm/(n+m)}r_{n,m} = o_P(1)$. Combined with (17), (32) and (33), this implies

$$\sqrt{\frac{nm}{n+m}}\left(\widehat{\overline{b}}_d - \overline{b}_d\right) = \frac{1}{E(\eta_d^2)}\left[\frac{\sqrt{\lambda}}{m^{1/2}}\sum_{i=1}^{m}(\psi_{1i} + \psi_{2i} + \psi_{3i}) + \frac{\sqrt{1-\lambda}}{n^{1/2}}\sum_{i=1}^{n}\psi_{4i}\right] + o_P(1).$$

The result then follows by $E[\psi_j] = 0$, $E(\psi_j^2) < \infty$ for all $j = 1, ..., 4$ and the central limit theorem.

Case 1: Assumption 3-(i) or (ii) holds. Let us assume that Assumption 3-(ii) holds. Then G^{-1} is continuous on $(0,1)$, and in particular on $F(\text{Supp}(Y))$. Let $\text{Supp}(Y) = \{y_1, ..., y_K\}$ and $y_0 = -\infty$. Remark that for any function q,

$$\int_0^1 F^{-1}(t)q(t)dt = \int_0^1 \sum_{k=1}^{K} y_k \mathbb{1}\left\{F(y_{k-1}) < t \leq F(y_k)\right\} q(t)dt,$$
$$\int_0^1 \widehat{F}^{-1}(t)q(t)dt = \int_0^1 \sum_{k=1}^{K} y_k \mathbb{1}\left\{\widehat{F}(y_{k-1}) < t \leq \widehat{F}(y_k)\right\} q(t)dt.$$

Let $I_j := (\widehat{F}(y_j), F(y_j)]$ if $\widehat{F}(y_j) < F(y_j)$, $I_j := (F(y_j), \widehat{F}(y_j)]$ otherwise. Then,

$$\begin{aligned}
&\int_0^1 (\widehat{F}^{-1}(t) - F^{-1}(t))q(t)dt \\
=&\int_0^1 \sum_{k=1}^{K} y_k \Big(\text{sgn}(\widehat{F}(y_k) - F(y_k)) \mathbb{1}\left\{t \in I_k\right\} - \text{sgn}(\widehat{F}(y_{k-1}) - F(y_{k-1})) \\
&\qquad \times \mathbb{1}\left\{t \in I_{k-1}\right\} \Big) q(t)dt \\
=&\int_0^1 \sum_{k=1}^{K-1} (y_k - y_{k+1}) \text{sgn}(\widehat{F}(y_k) - F(y_k)) \mathbb{1}\left\{t \in I_k\right\} q(t)dt. \qquad (34)
\end{aligned}$$

Now, we have, for $k = 1, ..., K-1$,

$$\begin{aligned}
\widehat{G}^{-1}\left(\min(\widehat{F}(y_k), F(y_k))\right) \leq& \frac{1}{|\widehat{F}(y_k) - F(y_k)|} \int_0^1 \mathbb{1}\left\{t \in I_k\right\} \widehat{G}^{-1}(t)dt \\
\leq& \widehat{G}^{-1}\left(\max(\widehat{F}(y_k), F(y_k))\right).
\end{aligned}$$

By continuity of G^{-1} at $F(y_k)$, $\widehat{G}^{-1}(F(y_k)) \xrightarrow{P} G^{-1}(F(y_k))$. Thus,

$$\frac{1}{|\widehat{F}(y_k) - F(y_k)|} \int_0^1 \mathbb{1}\left\{t \in I_k\right\} \widehat{G}^{-1}(t)dt \xrightarrow{P} G^{-1}(F(y_k)).$$

The same result holds replacing $\widehat{G}^{-1}$ by G^{-1}. Hence,

$$\frac{1}{|\widehat{F}(y_k) - F(y_k)|} \int_0^1 \mathbb{1}\left\{t \in I_k\right\} (\widehat{G}^{-1}(t) - G^{-1}(t))dt = o_P(1).$$

Then, replacing q by $\widehat{G}^{-1}(t) - G^{-1}(t)$ in (34), we obtain

$$\begin{aligned}
&\int_0^1 (\widehat{F}^{-1}(t) - F^{-1}(t))(\widehat{G}^{-1}(t) - G^{-1}(t))dt \\
=&\sum_{k=1}^{K-1} (y_k - y_{k+1}) \text{sgn}(\widehat{F}(y_k) - F(y_k)) \int_0^1 \mathbb{1}\left\{t \in I_k\right\} (\widehat{G}^{-1}(t) - G^{-1}(t))dt \\
=&\sum_{k=1}^{K-1} (y_k - y_{k+1})(\widehat{F}(y_k) - F(y_k)) \times o_P(1).
\end{aligned}$$

Then, because $\sqrt{nm/(n+m)}(\widehat{F}(y_k) - F(y_k)) = O_P(1)$, we obtain $R_{n,m} = o_P(1)$.

Now, if Assumption 3-(i) holds, the reasoning is the same as above, just reverting the roles of F and G, once we note that by assumption, F^{-1} is continuous at $G(\text{Supp}(\eta_d))$.

Case 2: Assumption 3-(iii) holds. We have, by Cauchy-Schwarz inequality,

$$R_{n,m}^2 \leq \frac{nm}{n+m} W_2^2(F_n, F) W_2^2(G_m, G).$$

Hence, by independence,

$$E\left[R_{n,m}^2\right] \leq \frac{nm}{n+m} E\left[W_2^2(F_n, F)\right] E\left[W_2^2(G_m, G)\right].$$

Now, assume that $Z = \eta_d$ in Assumption 3-(iii); the proof is the same if, instead, $Z = Y$. Theorem 1 in Fournier and Guillin (2015) shows that

$$E\left[W_2^2(F_n, F)\right] \lesssim n^{-1/2},$$

where "$\lesssim$" means that the inequality holds up to a number independent of (n, m). We now prove that

$$E\left[W_2^2(G_m, G)\right] = o(m^{-1/2}), \tag{35}$$

which implies that $E[R_{n,m}^2] = o(1)$ and concludes the proof by Markov inequality. First, remark that by Theorem 4.3 of Bobkov and Ledoux (2019),

$$E\left[W_2^2(G_m, G)\right] \leq \frac{2}{m} \sum_{i=1}^{m} V(\eta_{d(i)}), \tag{36}$$

where $\eta_{d(1)} < ... < \eta_{d(m)}$ denotes the order statistic of an i.i.d. sample $(\eta_{d1}, ..., \eta_{dm})$ from G. Then, by Condition (14) and Lemma 3, we have

$$\begin{aligned}
\sum_{i=1}^{m} V(\eta_{d(i)}) &\lesssim E\left[\sum_{i=1}^{m} \frac{1}{i \wedge (m+1-i)} \left(\frac{1}{C^2} \vee \frac{\eta_{d(i)}^2 \ln(1+|\eta_{d(i)}|)^4}{K^2} + \eta_{d(i)}^2\right)\right] \\
&\lesssim \left(E[Z_m^2] + E[Z_m^2 \ln(1+Z_m)^4]\right) \sum_{i=1}^{\left[\frac{m+1}{2}\right]} \frac{1}{i} \\
&\lesssim E[Z_m^{2+\varepsilon/3}] \left[1 + \ln(m)\right], \qquad (37)
\end{aligned}$$

where $Z_m = \max_{i=1,...,m}(|\eta_{di}|)$ and $[x]$ denotes the integer part of x. Now,

$$m^{-\frac{2+\varepsilon/3}{4+\varepsilon}} E[Z_m^{2+\varepsilon/3}] \leq \left\{m^{-1} E[Z_m^{4+\varepsilon}]\right\}^{\frac{2+\varepsilon/3}{4+\varepsilon}} = o(1), \tag{38}$$

where the inequality is due to Jensen's inequality and the equality holds by, e.g., Exercise 4 in Section 2.3 of van der Vaart and Wellner (2023) and because $E[|\eta_{d1}|^{4+\varepsilon}] < \infty$. Combining (36), (37) and (38), we obtain (35).

E.2 Theorem 4

Theorem 1 ensures that $\bar{b}_d > 0 > \bar{b}_{-d}$. Then, by Theorem 3, it suffices to prove the following:

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\psi}_{kj}^2 \xrightarrow{P} E[\psi_k^2], \quad \text{for } k \in \{1,...,3\}, \ \frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_{4i}^2 \xrightarrow{P} E[\psi_4^2] \tag{39}$$

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\psi}_{kj}\widehat{\psi}_{k'j} \xrightarrow{P} E[\psi_k\psi_{k'}] \ \text{ for } k,k' \in \{1,...,3\}, \ k \neq k'. \tag{40}$$

Hereafter, we let, for any $N \subset \mathbb{R}$ and $\varepsilon \geq 0$, $N^\varepsilon := \{x \in \mathbb{R} : \exists y \in N : |x-y| \leq \varepsilon\}$.

Eq. (39) holds for $k = 1$. We actually prove

$$\frac{1}{m}\sum_{j=1}^{m}(\widehat{\psi}_{1j} - \psi_{1j})^2 \xrightarrow{P} 0.$$

The result then follows by the triangle inequality and the LLN applied to the $(\psi_{1j}^2)_{j=1,...,m}$. By definition of $\widehat{\psi}_{1j}$ and ψ_{1j} and convexity of $x \mapsto x^2$,

$$(\widehat{\psi}_{1j} - \psi_{1j})^2 \leq 2\widehat{\bar{b}}_d^{\,2}\left(\widehat{\eta}_{dj}^2 - \frac{1}{m}\sum_{j=1}^{m}\widehat{\eta}_{dj}^2 - \eta_{dj}^2 + E[\eta_d^2]\right)^2 + 2(\eta_{dj}^2 - E[\eta_d^2])^2(\widehat{\bar{b}}_d - \bar{b}_d)^2.$$

The sample mean of the second term on the right-hand side converges to 0 in probability by Theorem 3 and $E[\eta_d^4] < \infty$. Recall that $\frac{1}{m}\sum_{j=1}^{m}\widehat{\eta}_{dj}^2 = \frac{1}{m}\sum_{j=1}^{m}\eta_{dj}^2 + o_P(m^{-1/2})$. Also, $\widehat{\bar{b}}_d = O_P(1)$. Then, using again convexity of $x \mapsto x^2$, it suffices to prove that

$$\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{\eta}_{dj}^2 - \eta_{dj}^2\right)^2 \xrightarrow{P} 0. \tag{41}$$

Remark that $(\widehat{\eta}_{dj}^2 - \eta_{dj}^2)^2 = (\widehat{\eta}_{dj} + \eta_{dj})^2(\widehat{\eta}_{dj} - \eta_{dj})^2$. Then, (41) follows from (55) in Lemma 2 and $(1/m)\sum_{j=1}^{m}(\widehat{\eta}_{dj} + \eta_{dj})^2 = O_P(1)$.

Eq. (39) holds for $k = 2$. Note that $\widehat{\psi}_{2j} = \widehat{\lambda}'\widehat{\delta}_j$, with $\widehat{\delta}_j = T_{-1j}\widehat{\eta}_{dj}$ and

$$\widehat{\lambda} = \left(\frac{1}{m}\sum_{j=1}^{m}T_{-1j}T_{-1j}'\right)^{-1}\left(\frac{1}{m}\sum_{j=1}^{m}\widehat{h}(\widehat{\eta}_{dj})T_{-1j}'\right).$$

This implies

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\psi}_{2j}^2 = \widehat{\lambda}'\left(\frac{1}{m}\sum_{j=1}^{m}\widehat{\delta}_j\widehat{\delta}_j'\right)\widehat{\lambda}.$$

It suffices to show convergence of $\widehat{\lambda}$ and the term in parentheses. Regarding the latter, we have

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\delta}_j\widehat{\delta}'_j = \frac{1}{m}\sum_{j=1}^{m}(T_{-1j}T'_{-1j})(\widehat{\eta}_{dj}-\eta_{dj})^2 + \frac{2}{m}\sum_{j=1}^{m}(T_{-1j}T'_{-1j})\eta_{dj}(\widehat{\eta}_{dj}-\eta_{dj}) + \frac{1}{m}\sum_{j=1}^{m}(T_{-1j}T'_{-1j})\eta_{dj}^2.$$

Moreover,

$$\left\|\frac{1}{m}\sum_{j=1}^{m}(T_{-1j}T'_{-1j})(\widehat{\eta}_{dj}-\eta_{dj})^2\right\| \leq \max_j(\widehat{\eta}_{dj}-\eta_{dj})^2\frac{1}{m}\sum_{j=1}^{m}\|T_{-1j}T'_{-1j}\|.$$

By Lemma 2, the left-hand side is an $o_P(1)$. Similarly, $(1/m)\sum_{j=1}^{m}(T_{-1j}T'_{-1j})\eta_{dj}(\widehat{\eta}_{dj}-\eta_{dj}) = o_P(1)$. Then, by the LLN,

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\delta}_j\widehat{\delta}'_j \xrightarrow{P} E[\eta_{dj}^2 T_{-1j}T'_{-1j}].$$

Let us turn to $\widehat{\lambda}$. It suffices to prove that

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{h}(\widehat{\eta}_{dj})T'_{-1j} \xrightarrow{P} E[h(\eta_d)T'_{-1}]. \tag{42}$$

Suppose first that $|\text{Supp}(\eta_d)| = |\text{Supp}(X)| < \infty$ and let $(u_1, ..., u_K) := \text{Supp}(\eta_d)$. Then, by (55), wpao, $|\{\widehat{\eta}_{d1}, ..., \widehat{\eta}_{dm}\}| = K$ and there exists a permutation σ such that both $\widehat{\eta}_{d\sigma(1)} \leq ... \leq \widehat{\eta}_{d\sigma(m)}$ and $\eta_{d\sigma(1)} \leq ... \leq \eta_{d\sigma(m)}$. If so,

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{h}(\widehat{\eta}_{dj})T'_{-1j} = \sum_{k=1}^{K}(G_n(u_k)-G_n(u_{k-1}))\overline{Y}_k\overline{T}'_{-1k},$$

with the convention that $u_0 = -\infty$ and where, letting $m_k := |\{j : \eta_{dj} = u_k\}|$, $\alpha_{n,m,k} := \lceil nG_m(u_{k-1})\rceil - 1$ and $\beta_{n,m,k} := \lceil nG_m(u_k)\rceil$,

$$\overline{T}_{-1k} := \frac{1}{m_k}\sum_{j:\eta_{dj}=u_k} T_{-1j},$$

$$\begin{aligned}\overline{Y}_k &:= \int_0^1 F_n^{-1}\left[G_m(u_{k-1}) + u(G_m(u_k)-G_m(u_{k-1}))\right]du \\ &= \frac{1}{\beta_{n,m,k}-\alpha_{n,m,k}}\left[\sum_{i=\alpha_{n,m,k}+1}^{\beta_{n,m,k}} Y_{(i)} - \left(\lambda^1_{n,m,k}Y_{(\alpha_{n,m,k}+1)} + \lambda^2_{n,m,k}Y_{(\beta_{n,m,k})}\right)\right],\end{aligned}$$

for some $(\lambda^1_{n,m,k}, \lambda^2_{n,m,k}) \in [0,1]^2$. By the LLN, $\overline{T}_{-1k} \xrightarrow{P} E[T_{-1}|\eta_d = u_k]$. Remark that $\alpha_{n,m,k}/n \xrightarrow{P} G(u_{k-1})$ and $\beta_{n,m,k}/n \xrightarrow{P} G(u_k)$. Then, by e.g. (22) p.681 in Shorack and Wellner (1986), we have

$$\frac{1}{\beta_{n,m,k}-\alpha_{n,m,k}}\sum_{i=\alpha_{n,m,k}+1}^{\beta_{n,m,k}} Y_{(i)} \xrightarrow{P} E\left[Y|Y \in [F^{-1}\circ G(u_{k-1}), F^{-1}\circ G(u_k)]\right] = h(u_k)$$

Moreover,

$$\frac{|Y_{(\beta_{n,m,k})}|}{\beta_{n,m,k}-\alpha_{n,m,k}} \leq \frac{\max_i |Y_i|/n}{G(u_k)-G(u_{k-1})+o_P(1)} \xrightarrow{P} 0.$$

and

$$\frac{|Y_{(\alpha_{n,m,k}+1)}|}{\beta_{n,m,k}-\alpha_{n,m,k}} \leq \frac{\max_i |Y_i|/n}{G(u_k)-G(u_{k-1})+o_P(1)} \xrightarrow{P} 0.$$

Hence, $\overline{Y}_k \xrightarrow{P} h(u_k)$. As a result,

$$\frac{1}{m}\sum_{j=1}^{m} \widehat{h}(\widehat{\eta}_{dj})T'_{-1j} \xrightarrow{P} \sum_{k=1}^{K}(G(u_k)-G(u_{k-1}))h(u_k)E[T'_{-1}|\eta_d=u_k] = E[h(\eta_d)T'_{-1}].$$

Next, let us prove (42) when G is continuous. By the LLN and Cauchy-Schwarz inequality, (42) holds if

$$\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{h}(\widehat{\eta}_{dj})-h(\eta_{dj})\right)^2 \xrightarrow{P} 0. \tag{43}$$

Fix $\delta > 0$. By Assumption 3 and the dominated convergence theorem, there exists $\overline{\varepsilon} > 0$ and a compact set $I \subset (0,1)$ such that (i) $E[\mathbb{1}\{G(\eta_d) \notin I\}] < \delta$; (ii) $I \subset I^{\overline{\varepsilon}} \subset (0,1)$; (iii) $I^{\overline{\varepsilon}} \cap \mathcal{D}_{F^{-1}} = \emptyset$, with $\mathcal{D}_{F^{-1}}$ the set of discontinuity points of F^{-1}. Since F^{-1} is continuous on $I^{\overline{\varepsilon}}$, it is also uniformly continuous on this compact set. Then, there exists $\varepsilon \in (0,\overline{\varepsilon})$ such that for all $(x,y) \in I^{\overline{\varepsilon}2}$, $|x-y| \leq \varepsilon$ implies $|F^{-1}(x)-F^{-1}(y)| \leq \delta^{1/2}$. Moreover, because $G(\eta_{dj}) \in I$ implies that η_{dj} belongs to a bounded set, by (56) and its variant in Lemma 2, wpao,

$$\max_{j:G(\eta_{dj})\in I} \max\left(|\widehat{G}(\widehat{\eta}_{dj})-G(\eta_{dj})|, |\widehat{G}(\widehat{\eta}^-_{dj})-G(\eta_{dj})|\right) \leq \varepsilon.$$

Then, under this event,

$$\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{h}(\widehat{\eta}_{dj})-h(\eta_{dj})\right)^2 \mathbb{1}\{G(\eta_{dj})\in I\} \qquad (44)\\
&\leq \max_{j:G(\eta_{dj})\in I}\int_0^1 |F_n^{-1}\left(\widehat{G}(\widehat{\eta}^-_{dj})+u(\widehat{G}(\widehat{\eta}_{dj})-\widehat{G}(\widehat{\eta}^-_{dj}))\right)-F^{-1}\circ G(\eta_{dj})|^2 du\\
&\leq \max_{j:G(\eta_{dj})\in I}\sup_{u\in[0,1]} |F_n^{-1}\left(\widehat{G}(\widehat{\eta}^-_{dj})+u(\widehat{G}(\widehat{\eta}_{dj})-\widehat{G}(\widehat{\eta}^-_{dj}))\right)-F^{-1}\circ G(\eta_{dj})|^2\\
&\leq 2\sup_{x\in I^\varepsilon}|F_n^{-1}(x)-F^{-1}(x)|^2 + 2\sup_{(t,u)\in I_\varepsilon^2:|t-u|\leq\varepsilon}|F^{-1}(t)-F^{-1}(u)|^2\\
&\leq 2\left[\sup_{x\in I^\varepsilon}|F_n^{-1}(x)-F^{-1}(x)|^2+\delta\right], \qquad (45)
\end{aligned}$$

where we have used Jensen's inequality for the first inequality. Since F^{-1} is continuous on $I^\varepsilon \subset (0,1)$, the first term on the right-hand side is smaller than δ wpao.

Then, to prove (43), it suffices to show that wpao,

$$\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{h}(\widehat{\eta}_{dj})-h(\eta_{dj})\right)^2 \mathbb{1}\left\{G(\eta_{dj})\notin I\right\}\leq q(\delta),$$

for some $q(\cdot)$ continuous at 0 and such that $q(0)=0$. By the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}&\left|\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{h}(\widehat{\eta}_{dj})-h(\eta_{dj})\right)^2 \mathbb{1}\left\{G(\eta_{dj})\notin I\right\}\right|\\ &\leq\left(\frac{1}{m}\sum_{j=1}^{m}\mathbb{1}\left\{G(\eta_{dj})\notin I\right\}\right)^{1/2}\left(\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{h}(\widehat{\eta}_{dj})-h(\eta_{dj})\right)^4\right)^{1/2}\end{aligned}$$

hence using that $\sum_{j=1}^{m}\mathbb{1}\left\{G(\eta_{dj})\notin I\right\}/m\xrightarrow{P}E[\mathbb{1}\left\{G(\eta_{dj})\notin I\right\}]<\delta$, it suffices to prove that

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{h}(\widehat{\eta}_{dj})^4=O_P(1),\quad \frac{1}{m}\sum_{j=1}^{m}h(\eta_{dj})^4=O_P(1).$$

The second result follows by the LLN, since $h(\eta_{dj})\overset{d}{=}Y_j$ and $E[Y^4]<\infty$. For the first, remark that by Jensen's inequality and since $\widehat{G}(\widehat{\eta}_{dj})-\widehat{G}(\widehat{\eta}_{dj}^{-})\geq 1/m$,

$$\begin{aligned}\widehat{h}(\widehat{\eta}_{dj})^4&\leq\int_0^1 F_n^{-1}[\widehat{G}(\widehat{\eta}_{dj}^{-})+u(\widehat{G}(\widehat{\eta}_{dj})-\widehat{G}(\widehat{\eta}_{dj}^{-}))]^4du\\ &\leq m\int_{\widehat{G}(\widehat{\eta}_{dj}^{-})}^{\widehat{G}(\widehat{\eta}_{dj})}F_n^{-1}(u)^4du.\end{aligned}$$

Hence, by Fubini-Tonelli's theorem,

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{h}(\widehat{\eta}_{dj})^4\leq\int_0^1 F_n^{-1}(u)^4du=\frac{1}{n}\sum_{i=1}^{n}Y_i^4=O_P(1).$$

The result follows.

Eq. (39) holds for $k=3$. Let $a\in(0,1)$ be a continuity point of G^{-1} and F^{-1}. We have

$$\begin{aligned}\psi_3&=-\int\left[1\{\eta_d\leq u\}-G(u)\right]F^{-1}\circ G(u)du\\ &=\int_{G^{-1}(1)}^{G^{-1}(\xi_d)}F^{-1}\circ G(u)du+\int G(u)F^{-1}\circ G(u)du\\ &=\int_{G^{-1}(a)}^{G^{-1}(\xi_d)}F^{-1}\circ G(u)du+\int G(u)F^{-1}\circ G(u)du-\int_{G^{-1}(a)}^{G^{-1}(1)}F^{-1}\circ G(u)du,\end{aligned}$$

for some $\xi_d\sim\mathcal{U}[0,1]$. As a result,

$$E[\psi_3]=\int_0^1\int_{G^{-1}(a)}^{G^{-1}(t)}F^{-1}(G(s))dsdt+\int G(u)F^{-1}\circ G(u)du-\int_{G^{-1}(a)}^{G^{-1}(1)}F^{-1}\circ G(u)du.$$

Since, by Fubini's theorem, we also have $E[\psi_3] = 0$, we obtain

$$\psi_3 = \psi_3 - E[\psi_3] = \overline{c}_2(\xi_d, F, G), \tag{46}$$

where

$$\begin{aligned} \overline{c}_2(t, F, G) :=& c_2(t, F, G) - \int_0^1 c_2(u, F, G)du \\ c_2(t, F, G) :=& \int_{G^{-1}(a)}^{G^{-1}(t)} F^{-1}(G(s))ds, \end{aligned} \tag{47}$$

Moreover, using that $\widehat{G}_m^{-1}(t) = \widehat{\eta}_{d,(i)}$ for $t \in ((i-1)/m, i/m]$ and all $i = 1, \ldots, m$,

$$\frac{1}{m}\sum_{j=1}^{m} \widehat{\psi}_{3j}^2 = \int_0^1 \overline{c}_2^2(t, F_n, \widehat{G}_m)dt. \tag{48}$$

By Lemma 4 below and the continuous mapping theorem, it suffices to show that $\mathcal{W}_4(\widehat{G}_m, G) \xrightarrow{P} 0$ and either $\mathcal{W}_4(F_n, F) \xrightarrow{P} 0$ (if Assumption 3-(ii) or (iii) holds), or $\mathcal{W}_2(F_n, F) \xrightarrow{P} 0$ (if Assumption 3-(i) holds). The condition on F_n holds by, e.g., Theorem 2.13 in Bobkov and Ledoux (2019). By the same theorem, $\mathcal{W}_4(G_m, G)$ converge to 0 a.s. Moreover,

$$\mathcal{W}_4(\widehat{G}_m, G_m) \leq \left[\frac{1}{m}\sum_{i=1}^{m}(\eta_{di} - \widehat{\eta}_{di})^4\right]^{1/4} \leq \max_{i=1,\ldots,m} |\eta_{di} - \widehat{\eta}_{di}| \xrightarrow{P} 0,$$

where the first inequality follows by definition of $\mathcal{W}_4$ and the convergence holds by Lemma 2. Then, $\mathcal{W}_4(\widehat{G}_m, G) \xrightarrow{P} 0$ follows by the triangle inequality.

Eq. (39) holds for $k = 4$. The reasoning is the same as for $k = 3$: we just exchange the roles of F and G and note that Lemma 4 still applies then.

Eq. (40) holds for $(k, k') = (1, 2)$. We have

$$\widehat{\psi}_{1j}\widehat{\psi}_{2j} = -\widehat{\overline{b}}_d \left(\widehat{\eta}_{dj}^2 - \frac{1}{m}\sum_{k=1}^{m} \widehat{\eta}_{dk}^2\right) \widehat{\lambda}' T_{-1j} \widehat{\eta}_{dj}.$$

The result follows by convergences of $\widehat{\lambda}$ and $\widehat{\overline{b}}_d$, and Eq. (55) in Lemma 2.

Eq. (40) holds for $(k, k') = (1, 3)$. It suffices to prove that

$$\frac{1}{m}\sum_{j=1}^{m} \widehat{\eta}_{dj}^2 \widehat{\psi}_{3j} \xrightarrow{P} E[\eta_d^2 \psi_3].$$

To this end, note that by (46)

$$\eta_d^2\psi_3 = (G^{-1}(\xi_d))^2\overline{c}_2(\xi_d, F, G).$$

Hence by the same argument as (48),

$$\frac{1}{m}\sum_{j=1}^{m}\widehat{\eta}_{dj}^2\widehat{\psi}_{3j} = \int_0^1 \overline{\widetilde{c}}_2(t, F_n, \widehat{G}_m)dt,$$

where $\overline{\widetilde{c}}_2(t,F,G) := \widetilde{c}_2(t,F,G) - \int_0^1 (G^{-1}(u))^2\,du \int_0^1 c_2(u,F,G)du$ and

$$\widetilde{c}_k(t,F,G) := \left(G^{-1}(t)\right)^k \int_{G^{-1}(a)}^{G^{-1}(t)} F^{-1}(G(s))ds.$$

We obtain the result using Lemma 4 below, the continuous mapping theorem, and the fact as shown above, we have $\mathcal{W}_4(\widehat{G}_m, G) \xrightarrow{P} 0$ and either $\mathcal{W}_4(F_n, F) \xrightarrow{P} 0$ (if Assumption 3-(ii) or (iii) holds), or $\mathcal{W}_2(F_n, F) \xrightarrow{P} 0$ (if Assumption 3-(i) holds).

Eq. (40) **holds for** $(k, k') = (2, 3)$**.** We have

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^{m}\widehat{\psi}_{2j}\widehat{\psi}_{3j} &= -\frac{1}{m}\sum_{j=1}^{m}\widehat{\lambda}'T_{-1j}\widehat{\eta}_{dj}\widehat{\psi}_{3j} \\ &= -\frac{1}{m}\sum_{j=1}^{m}\lambda'T_{-1j}\widehat{\eta}_{dj}\widehat{\psi}_{3j} + o_p(1),\end{aligned}$$

using the convergence of $\widehat{\lambda}$. Then,

$$-\frac{1}{m}\sum_{j=1}^{m}\lambda'T_{-1j}\widehat{\eta}_{dj}\widehat{\psi}_{3j} = -\frac{1}{m}\sum_{j=1}^{m}\lambda'T_{-1j}(\widehat{\eta}_{dj}\widehat{\psi}_{3j} - \eta_{dj}\psi_{3j}) - \frac{1}{m}\sum_{j=1}^{m}\psi_{2j}\psi_{3j}.$$

Using the Cauchy-Schwarz inequality, we have

$$\left|\frac{1}{m}\sum_{j=1}^{m}\lambda'T_{-1j}(\widehat{\eta}_{dj}\widehat{\psi}_{3j} - \eta_{dj}\psi_{3j})\right| \le \left(\frac{1}{m}\sum_{j=1}^{m}(\lambda'T_{-1j})^2\right)^{1/2}\left(\frac{1}{m}\sum_{j=1}^{m}(\widehat{\eta}_{dj}\widehat{\psi}_{3j} - \eta_{dj}\psi_{3j})^2\right)^{1/2}$$

The first term on the right hand side is bounded in probability by the LLN, since $E[\|X\|^4] < \infty$. By the LLN for the term $\frac{1}{m}\sum_{j=1}^{m}\psi_{2j}\psi_{3j}$ it suffices to show that $\frac{1}{m}\sum_{j=1}^{m}(\widehat{\eta}_{dj}\widehat{\psi}_{3j} - \eta_{dj}\psi_{3j})^2 \xrightarrow{P} 0$. We have, using $\eta_d\psi_3 = \widetilde{c}_1(\xi_d, F, G)$ with $\xi_d \sim \mathcal{U}[0,1]$, that

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}(\widehat{\eta}_{di}\widehat{\psi}_{3i} - \eta_{di}\psi_{3i})^2 &= \sum_{i=1}^{m}\int_{(i-1)/m}^{i/m}\left(\widetilde{c}_1(t, F_n, \widehat{G}_m) - \widetilde{c}_1(\xi_{d,(i)}, F, G)\right)^2 dt \\ &\le 2\sum_{i=1}^{m}\int_{(i-1)/m}^{i/m}\left(\widetilde{c}_1(t, F_n, \widehat{G}_m) - \widetilde{c}_1(t, F, G)\right)^2 dt\end{aligned}$$

$$
\begin{aligned}
&+ 2\sum_{i=1}^{m}\int_{(i-1)/m}^{i/m}\left(\widetilde{c}_1(t,F,G)-\widetilde{c}_1(\xi_{d,(i)},F,G)\right)^2 dt\\
&\leq 2\int_0^1\left(\widetilde{c}_1(t,F_n,\widehat{G}_m)-\widetilde{c}_1(t,F,G)\right)^2 dt\\
&\quad + 2\sum_{i=1}^{m}\int_{(i-1)/m}^{i/m}\left(\widetilde{c}_1(t,F,G)-\widetilde{c}_1(\xi_{d,(i)},F,G)\right)^2 dt.
\end{aligned}
$$

By Lemma 4 and the continuous mapping theorem, $\int_0^1\left(\widetilde{c}_1(t,F_n,\widehat{G}_m)-\widetilde{c}_1(t,F,G)\right)^2 dt \xrightarrow{P} 0$. Next, let us prove that

$$
\Delta := \sum_{i=1}^{m}\int_{(i-1)/m}^{i/m}\left(\widetilde{c}_1(t,F,G)-\widetilde{c}_1(\xi_{d,(i)},F,G)\right)^2 dt \xrightarrow{P} 0. \tag{49}
$$

Let $F_{\xi,n}$ denote the empirical cdf of the $(\xi_{d,i})_{i=1,\ldots,m}$. We have (see, e.g. Shorack and Wellner, 1986, Eq. (11) p.86)

$$
\max_{i=1,\ldots,m}\left|\xi_{d,(i)}-\frac{i}{m}\right| \leq \sup_{t\in[0,1]}|F_{\xi,n}^{-1}(t)-t| \xrightarrow{P} 0. \tag{50}
$$

Fix $\varepsilon>0$ and $M>1$. Because continuous functions are dense in $L^2([0,1])$, there exists a continuous function $\widetilde{c}_1^c$ such that

$$
\Delta_1 := \int_0^1\left(\widetilde{c}_1(t,F,G)-\widetilde{c}_1^c(t)\right)^2 dt < \frac{\varepsilon}{6M}. \tag{51}
$$

Moreover,

$$
\Delta \leq 3(\Delta_1+\Delta_2+\Delta_3), \tag{52}
$$

where

$$
\begin{aligned}
\Delta_2 :=& \sum_{i=1}^{m}\int_{(i-1)/m}^{i/m}\left(\widetilde{c}_1^c(t)-\widetilde{c}_1^c(\xi_{d,(i)})\right)^2 dt,\\
\Delta_3 :=& \frac{1}{m}\sum_{i=1}^{m}\left(\widetilde{c}_1^c(\xi_{d,i})-\widetilde{c}_1(\xi_{d,i},F,G)\right)^2.
\end{aligned}
$$

Since $\widetilde{c}_1^c$ is uniformly continuous on $[0,1]$, there exists $\delta>0$ such that for all $(x,y)\in K^2$, $|x-y|\leq\delta$ implies that $|\widetilde{c}_1(x)-\widetilde{c}_1(y)|\leq[\varepsilon/(6M)]^{1/2}$. Combined with (50), this implies that for all $m>2/\delta$,

$$
P\left(\Delta_2>\frac{\varepsilon}{6M}\right)\leq\frac{1}{M}. \tag{53}
$$

Finally, by Markov's inequality, for all $q>0$,

$$
P\left(\Delta_3>q\varepsilon\right)\leq\frac{E[\Delta_3]}{q\varepsilon}=\frac{\Delta_1}{q\varepsilon}<\frac{1}{q\,6M}. \tag{54}
$$

Using (51)-(54), we finally obtain

$$\begin{aligned} P(\Delta > \varepsilon) \leq & P(\Delta_2 > \varepsilon/(6M)) + P\left(\Delta_2 \leq \varepsilon/(6M), \Delta_3 > \varepsilon(1-1/(2M))/3\right) \\ \leq & \frac{1}{M} + \frac{1}{((1-1/(2M))\,2M} = \frac{1}{M} + \frac{1}{2M-1}. \end{aligned}$$

Eq. (49) follows since $\varepsilon > 0$ and $M > 1$ were arbitrary.

E.3 Additional lemmas

The proof of Theorem 3 relies on two lemmas, which we state and prove below. Note that Lemma 3 is similar to Corollary 2.12 in Boucheron and Thomas (2015) but handles variables taking negative values. Also, Lemma 4 is similar to Lemma A.1 in Del Barrio et al. (2019) but holds under slightly weaker conditions.

Lemma 1 *For any cdfs F, G, $Y = F^{-1}(U)$ and $U \sim \mathcal{U}[0,1]$, we have*

$$\int_0^1 [\mathbb{1}\{U \leq t\} - t] G^{-1}(t) dF^{-1}(t) = \int_{-\infty}^{\infty} [\mathbb{1}\{Y \leq u\} - F(u)] G^{-1} \circ F(u) du.$$

Lemma 2 *Suppose Assumptions 3 hold. Then, for all $C > 0$ and $x \in \mathbb{R}$ continuity point of G,*

$$\max_{j \in \{1,\ldots,m\}} |\widehat{\eta}_{dj} - \eta_{dj}| \xrightarrow{P} 0, \tag{55}$$

$$\max_{j : |\eta_{dj}| \leq C} |\widehat{G}(\widehat{\eta}_{dj}) - G(\eta_{dj})| \xrightarrow{P} 0, \tag{56}$$

$$\widehat{G}(x) \xrightarrow{P} G(x). \tag{57}$$

Moreover, if G is continuous, (56) *still holds if we replace $\widehat{G}(\widehat{\eta}_{dj})$ by $\widehat{G}(\widehat{\eta}_{dj}^{-})$.*

Lemma 3 *Suppose that T has a cdf F, survival function S and a positive density f. Then, for all $i \in \{1, ..., n\}$,*

$$V(T_{(i)}) \leq \frac{32}{i \wedge (n+1-i)} E\left[2\left(\frac{F(T_{(i)})S(T_{(i)})}{f(T_{(i)})}\right)^2 + T_{(i)}^2\right].$$

Lemma 4 *For $f \in \{\bar{c}_2, \widetilde{c}_1, \widetilde{c}_2\}$, the function $(F, G) \mapsto f(\cdot, F, G)$ is continuous in $L^2[0,1]$ ($L^1[0,1]$ for $f = \widetilde{c}_2$) with respect to the metric*

$$d[(F,G),(F',G')] = \mathcal{W}_4(F', F) + \mathcal{W}_4(G', G),$$

at any (F_0, G_0) (and also at (G_0, F_0) when $f = \bar{c}_2$) satisfying the same restrictions as (F, G) in Assumption 3-(ii) or (iii). The same holds if in the metric d, we replace $\mathcal{W}_4(F', F)$ by $\mathcal{W}_2(F', F)$, provided that (F_0, G_0) satisfy the same restrictions as (F, G) in Assumption 3-(i).

E.3.1 Proof of Lemma 1

Note that F is a generalized inverse of F^{-1} (see, e.g., Shorack and Wellner, 1986, p.7). Then, by, e.g., Eq. (1) in Falkner and Teschl (2012),

$$\int_0^1 [\mathbb{1}\{U \leq t\} - t] G^{-1}(t) dF^{-1}(t) = \int_{-\infty}^{\infty} [\mathbb{1}\{U \leq F(u)\} - F(u)] G^{-1} \circ F(u) du.$$

The result follows by noting that $U \leq F(u)$ if and only if $Y \leq u$ (see, e.g., Lemma 21.1 in van der Vaart, 2000).

E.3.2 Proof of Lemma 2

First,

$$\begin{aligned}
\max_{i=1,\ldots,m} |\widehat{\eta}_{di} - \eta_{di}| &= \max_{i=1,\ldots,m} \left| T'_{-1i}(\widehat{\gamma} - \gamma_0) \right| \\
&\leq \left[\max_{i=1,\ldots,m} \|T_{-1i}\| \right] \|\widehat{\gamma} - \gamma_0\| \\
&= o_P(m^{1/2}) \times O_P(m^{-1/2}) \\
&= o_P(1).
\end{aligned}$$

The second equality follows since $E[\|T_{-1i}\|^2] < \infty$, see e.g. Exercise 4 in Section 2.3 of van der Vaart and Wellner (2023).

Let us turn to (56). First, assume that (i) holds in Assumption 3. Given that $|\text{Supp}(\eta_d)| < \infty$, it suffices to prove that for all $u \in \text{Supp}(\eta_d)$,

$$\max_{j:\eta_{dj}=u} \left| \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\widehat{\eta}_{dk} = \widehat{\eta}_{dj}\} - P(\eta_d = u) \right| \xrightarrow{P} 0. \tag{58}$$

Fix such a $u \in \text{Supp}(\eta_d)$. Since $|\text{Supp}(\eta_d)| = |\text{Supp}(X)|$, there exists $x \in \text{Supp}(X)$ such that for all j, $\eta_{dj} = u \Leftrightarrow X_j = x$. Also, because of (55), there exists $u_m \xrightarrow{P} u$ such that $X_j = x \Leftrightarrow \widehat{\eta}_{dj} = u_m$. Hence, for sufficiently large m,

$$\frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\widehat{\eta}_{dk} = \widehat{\eta}_{dj}\} = \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}\{\eta_{dk} = \eta_{dj}\}.$$

The result follows from the law of large numbers.

Now, assume that (ii) or (iii) holds in Assumption 3. Fix $\delta > 0$. Since G is continuous, there exists $\varepsilon > 0$ such that for all $(x, y) \in [-C - 2\varepsilon, C + 2\varepsilon]^2$, $|x - y| \leq 2\varepsilon$ implies $|G(y) - G(x)| < \delta$. By (55), with probability approaching

one, $|\widehat{\eta}_{dj} - \eta_{dj}| \leq \varepsilon$ for all j. Under this event, $\widehat{G}(x) \in [G_m(x-\varepsilon), G_m(x+\varepsilon)]$ for all $x \in \mathbb{R}$. Thus,

$$\begin{aligned}
&\max_{j:|\eta_{dj}|\leq C} |\widehat{G}(\widehat{\eta}_{dj}) - G(\eta_{dj})| \\
\leq &\max_{j:|\eta_{dj}|\leq C} \max\left\{|G_m(\eta_{dj}+2\varepsilon) - G(\eta_{dj})|, |G_m(\eta_{dj}-2\varepsilon) - G(\eta_{dj})|\right\} \\
\leq &\sup_{x\in[-C-2\varepsilon, C+2\varepsilon]} |G_m(x) - G(x)| + \sup_{(x,y)\in[-C-2\varepsilon,C+2\varepsilon]^2:|x-y|\leq 2\varepsilon} |G(x) - G(y)| \\
\leq &\sup_{x\in\mathbb{R}} |G_m(x) - G(x)| + \delta.
\end{aligned}$$

By Glivenko-Cantelli theorem, $\sup_{x\in\mathbb{R}} |G_m(x) - G(x)| < \delta$ with probability approaching one. Eq. (56) follows since $\delta > 0$ was arbitrary. To see that Eq. (56) still holds with $\widehat{G}(\widehat{\eta}_{dj})$ replaced by $\widehat{G}(\widehat{\eta}^-_{dj})$, remark that there exists $(\widetilde{\eta}_{dj})_{j=1,\ldots,m}$ such that $\widehat{G}(\widetilde{\eta}_{dj}) = \widehat{G}(\widehat{\eta}^-_{dj})$ and $|\widetilde{\eta}_{dj} - \eta_{dj}| \leq \varepsilon/2$. Then, the same proof as above applies, once we remark that with probability approaching one, $|\widehat{\eta}_{dj} - \eta_{dj}| \leq \varepsilon/2$ for all j.

Finally, we prove (57). If (i) holds in Assumption 3, a continuity point x of G is such that either $x < \min(\text{Supp}(\eta_d))$, $x > \max(\text{Supp}(\eta_d))$ or there exists $(u, v) \in \text{Supp}(\eta_d)^2$ such that $u < x < v$ (and $G(x) = G(u)$). In the first case, wpao, $x < \min_j \widehat{\eta}_j$ and thus $\widehat{G}(x) = 0$. The reasoning with $x > \max(\text{Supp}(\eta_d))$ is the same. In the third case, wpao,

$$\underline{x} := \max\{\widehat{\eta}_{dj} : \eta_{dj} = u\} < x < \overline{x} := \min\{\widehat{\eta}_{dj} : \eta_{dj} = v\}$$

and there is no j such that $\widehat{\eta}_{dj} \in (\underline{x}, \overline{x})$. Hence, under this event,

$$\widehat{G}\left(\max\{\widehat{\eta}_{dj} : \eta_{dj} = u\}\right) = \widehat{G}(x).$$

By (56), the left-hand side converges in probability to $G(u)$. The result follows.

Now, assume that (ii) or (iii) holds in Assumption 3. Fix $\delta > 0$ and let $\varepsilon > 0$ be such that $|y - x| < \varepsilon$ implies $|G(y) - G(x)| < \delta$. We proved above that $\widehat{G}(x) \in [G_m(x-\varepsilon), G_m(x+\varepsilon)]$ wpao. By the law of large numbers, $G_m(x-\varepsilon) \xrightarrow{P} G(x-\varepsilon)$ and $G_m(x+\varepsilon) \xrightarrow{P} G(x+\varepsilon)$. Hence, wpao, $|\widehat{G}(x) - G(x)| < 2\delta$. The result follows since $\delta > 0$ was arbitrary.

E.3.3 Proof of Lemma 3

First, note that

$$V(T_{(i)}) \leq 2\left[V(T_{(i)}F(T_{(i)})) + V(T_{(i)}S(T_{(i)}))\right]. \tag{59}$$

Remark that $T_{(i)} = F^{-1}(1 - \exp(-E_{(i)}))$, where $(E_1, ..., E_n)$ are iid, Exponential variables of parameter 1. Then, by Rényi's representation of order statistics for such variables,

$$V(T_{(i)}F(T_{(i)})) = V\left[F^{-1}\left(1 - e^{-\sum_{k=n+1-i}^{n} E_k/k}\right)\left(1 - e^{-\sum_{k=n+1-i}^{n} E_k/k}\right)\right].$$

Let us define

$$g(x_{n+1-i}, ..., x_n) = F^{-1}\left(1 - e^{-\sum_{k=n+1-i}^{n} x_k/k}\right)\left(1 - e^{-\sum_{k=n+1-i}^{n} x_k/k}\right).$$

Then, by Poincare's inequality for exponential variables (see, e.g., Proposition 2.10 in Boucheron and Thomas, 2015), we have

$$V(T_{(i)}F(T_{(i)})) \leq 4E\left[\sum_{k=n+1-i}^{n} \frac{\partial g}{\partial x_k}(E_{n+1-i}, ..., E_n)^2\right].$$

Remark that for all $j \in \{n+1-i, ..., n\}$,

$$\begin{aligned}\frac{\partial g}{\partial x_j}(x_{n+1-i}, ..., x_n) = \frac{1}{j}\Bigg[&\frac{1 - e^{-\sum_{k=n+1-i}^{n} x_k/k}}{h \circ F^{-1}\left(1 - e^{-\sum_{k=n+1-i}^{n} x_k/k}\right)} \\ &+ e^{-\sum_{k=n+1-i}^{n} x_k/k} F^{-1}\left(1 - e^{-\sum_{k=n+1-i}^{n} x_k/k}\right)\Bigg].\end{aligned}$$

Thus,

$$\begin{aligned}V(T_{(i)}F(T_{(i)})) \leq& 4E\left[\sum_{k=n+1-i}^{n} \frac{\partial g}{\partial x_k}(E_{n+1-i}, ..., E_n)^2\right] \\ =& 4\left[\sum_{j=n+1-i}^{n} \frac{1}{j^2}\right] E\left[\left(\frac{F(T_{(i)})}{h(T_{(i)})} + S(T_{(i)})T_{(i)}\right)^2\right] \\ \leq& \frac{16}{n+1-i} E\left[\left(\frac{F(T_{(i)})S(T_{(i)})}{f(T_{(i)})}\right)^2 + S(T_{(i)})^2 T_{(i)}^2\right]. \qquad (60)\end{aligned}$$

To deal with $V(T_{(i)}S(T_{(i)}))$, we use $T_{(i)} = F^{-1}(\exp(-E_{(n+1-i)}))$ and reason exactly as above. This yields:

$$V(T_{(i)}S(T_{(i)})) \leq \frac{16}{i} E\left[\left(\frac{F(T_{(i)})S(T_{(i)})}{f(T_{(i)})}\right)^2 + F(T_{(i)})^2 T_{(i)}^2\right]. \qquad (61)$$

By combining (59), (60), (61) and $x^2 + (1-x)^2 \leq 1$ for $0 \leq x \leq 1$, we finally obtain

$$V(T_{(i)}) \leq \frac{32}{i \wedge (n+1-i)} E\left[2\left(\frac{F(T_{(i)})S(T_{(i)})}{f(T_{(i)})}\right)^2 + T_{(i)}^2\right].$$

E.3.4 Proof of Lemma 4

We mostly focus on $f = \bar{c}_2^2$, as the reasoning is the same when $f \in \{\widetilde{c}_1, \widetilde{c}_2\}$. As in (46), we assume without loss of generality that a is a continuity point of G_0^{-1} and F_0^{-1}. Consider a sequence $(F_n, G_n)_{n\geq 1}$ converging to (F_0, G_0) satisfying Assumption 3-(ii) or (iii). We first show that $c_2(t, F_n, G_n) \to c_2(t, F_0, G_0)$ for almost every $t \in (0, 1)$. Then, we prove that $c_2(\cdot, F_n, G_n) \to c_2(\cdot, F, G)$ in $L_2[0, 1]$. The continuity result at (F_0, G_0) follows. Then, we show how to adapt the reasoning to prove continuity at (G_0, F_0) instead of (F_0, G_0). Next, we prove the result with the alternative metric when (F_0, G_0) satisfy Assumption 3-(i). Finally, we show how to adapt the argument when $f \in \{\widetilde{c}_1, \widetilde{c}_2\}$.

1. $c_2(t, F_n, G_n) \to c_2(t, F_0, G_0)$ **for almost every** $t \in (0, 1)$**.** Assume without loss of generality that $t \in (a, 1)$, and that it is a continuity point of G_0^{-1}. Suppose also that s (a) is a continuity point of G_0; (b) is such that $G_0(s)$ is a continuity point of F_0^{-1}; (c) satisfies $s \notin \{G_0^{-1}(a), G_0^{-1}(t)\}$. Note that by, e.g., Theorem 6.9 of Villani (2009) and Lemma 21.2 in van der Vaart (2000), $G_n(s') \to G_0(s')$ for all continuity points s' of G_0, and $F_n^{-1}(u) \to F_0^{-1}(u)$ for all continuity points of F_0. Fix $\delta > 0$ and let $\varepsilon > 0$ be such that $|u - G_0(s)| \leq \varepsilon$ implies $|F_0^{-1}(u) - F_0^{-1}(G_0(s))| \leq \delta$ and $G_0(s) - \varepsilon$ and $G_0(s) + \varepsilon$ are continuity points of F_0^{-1}. Such an ε exists since the set of discontinuity points of F_0^{-1} is at most countable. Now, for all n large enough, $|G_n(s) - G_0(s)| \leq \varepsilon$, which implies by monotonicity that $F_n^{-1}(G_n(s)) \in [F_n^{-1}(G_0(s) - \varepsilon), F_n^{-1}(G_0(s) + \varepsilon)]$. Then, by construction,

$$F_n^{-1}(G_0(s) - \varepsilon) \to F_0^{-1}(G_0(s) - \varepsilon) \geq F_0^{-1}(G_0(s)) - \delta,$$

and similarly for $F_n^{-1}(G_0(s) + \varepsilon)$. Since δ was arbitrary, we obtain $F_n^{-1}(G_n(s)) \to F_0^{-1}(G_0(s))$. In turn, because $s \notin \{G_0^{-1}(a), G_0^{-1}(t)\}$,

$$F_n^{-1}(G_n(s))\mathbb{1}\left\{G_n^{-1}(a) \leq s \leq G_n^{-1}(t)\right\} \to F_0^{-1}(G_0(s))\mathbb{1}\left\{G_0^{-1}(a) \leq s \leq G_0^{-1}(t)\right\}. \quad (62)$$

Now remark that under Assumption 3-(ii) or (iii), Conditions (a), (b) and (c) above hold for almost every s. Hence, (62) holds for almost all s. Moreover, for all n large enough, $s \mapsto |F_n^{-1}(G_n(s))|\mathbb{1}\{G_n^{-1}(a) \leq s \leq G_n^{-1}(t)\}$ is bounded by some $K > 0$. Then, by the dominated convergence theorem, $c_2(t, F_n, G_n) \to c_2(t, F_0, G_0)$. Since almost all t are continuity point of G_0^{-1}, Point 1 follows.

2. $c_2(\cdot, F_n, G_n) \to c_2(\cdot, F, G)$ **in** $L_2[0,1]$**.** Given Step 1, it suffices to prove, by Lebesgue–Vitali theorem (see e.g., Theorem 4.5.4 in Bogachev, 2007), that

$$\lim_{M\to\infty} \sup_{n\geq 1} \int_0^1 c_2(t, F_n, G_n)^2 \mathbb{1}\left\{c_2(t, F_n, G_n)^2 > M\right\} dt = 0. \tag{63}$$

First, remark that for all $s < G_n^{-1}(t)$, $G_n(s) \leq t$. Hence, for such s, $F_n^{-1}(G_n(s)) \leq F_n^{-1}(t)$. Similarly, for $s \geq G_n^{-1}(a)$, $F_n^{-1}(G_n(s)) \geq F_n^{-1}(a)$. As a result,

$$\begin{aligned} |c_2(t, F_n, G_n)| \leq & \left|G_n^{-1}(t) - G_n^{-1}(a)\right| \left(\left|F_n^{-1}(t)\right| + \left|F_n^{-1}(a)\right|\right) \\ \leq & q(G_n, t) \times q(F_n, t), \end{aligned} \tag{64}$$

where $q(F,t) := |F^{-1}(t)| + |F^{-1}(a)|$ for any cdf F. Then,

$$\mathbb{1}\left\{c_2(t, F_n, G_n)^2 > M\right\} \leq \mathbb{1}\left\{q(G_n, t) > \sqrt{M}\right\} + \mathbb{1}\left\{q(F_n, t) > \sqrt{M}\right\}.$$

As a result, by Cauchy-Schwarz inequality,

$$\begin{aligned} & \int_0^1 c_2(t, F_n, G_n)^2 \mathbb{1}\left\{c_2(t, F_n, G_n)^2 > M\right\} dt \\ \leq & \left[\int_0^1 q(F_n, t)^4 dt \int_0^1 q(G_n, t)^4 \mathbb{1}\left\{q(G_n, t) > \sqrt{M}\right\} dt\right]^{1/2} \\ & + \left[\int_0^1 q(F_n, t)^4 \mathbb{1}\left\{q(F_n, t) > \sqrt{M}\right\} dt \int_0^1 q(G_n, t)^4 dt\right]^{1/2}. \end{aligned}$$

We have $G_n^{-1}(a) \to G_0^{-1}(a)$. Also, because $\mathcal{W}_4(G_n, G_0) \to 0$, $G_n^{-1} \to G^{-1}$ in $L_4[0,1]$. Then, we also have $q(G_n, \cdot) \to q(G_0, \cdot)$ in $L_4[0,1]$. By Lebesgue–Vitali theorem again, $\lim_{M\to\infty} \sup_{n\geq 1} \int_0^1 q(G_n, t)^4 \mathbb{1}\{q(G_n, t) > \sqrt{M}\} dt = 0$. The same holds with F_n instead of G_n. Equation (63) follows.

3. Continuity at (G_0, F_0)**.** The reasoning is the same as above. Step 2 holds as is. In Step 1, we just need to check that Condition (b) on s still holds for almost every s when exchanging the roles of F_0 and G_0. This is true under Assumption 3-(ii), since G_0^{-1} is continuous on $(0,1)$. This also holds under Assumption 3-(iii): F_0 is strictly increasing on its support (since F_0^{-1} is continuous), and the set of discontinuity points of G_0^{-1} is at most countable.

4. Continuity with respect to the alternative metric. The proof is very similar as above. Step 1 is as above, once we note that Condition (b) still holds for almost every s under Assumption 3-(i). Regarding Step 2, start from (64) and use instead that $q(G_n, t)$ is bounded for $t \in (0,1)$. Since $q(F_n, \cdot)^2$ is uniformly integrable, $c_2(\cdot, F_n, G_n)^2$ is uniformly integrable as well. The result follows.

5. Adaptation to $f \in \{\widetilde{c}_1, \widetilde{c}_2\}$. The reasoning in Part 1 above is the same. For Part 2, let us just consider $f = \widetilde{c}_2$; with $f = \widetilde{c}_1$, we simply have to adjust some exponents and the use of Hölder's inequality below. The reasoning is the same as above but we use instead the inequality $|\widetilde{c}_2(t, F_n, G_n)| \leq q(G_n, t)^3 \times q(F_n, t)$. Then,

$$\mathbb{1}\left\{\widetilde{c}_2(t, F_n, G_n) > M\right\} \leq \mathbb{1}\left\{q(G_n, t) > M^{1/3}\right\} + \mathbb{1}\left\{q(F_n, t) > M\right\}.$$

As a result, by Hölder's inequality with exponents $4/3$ and 4,

$$\begin{aligned}
&\int_0^1 \widetilde{c}_2(t, F_n, G_n)\mathbb{1}\left\{\widetilde{c}_2(t, F_n, G_n) > M\right\} dt \\
&\leq \left[\int_0^1 q(F_n, t)^4 dt \int_0^1 q(G_n, t)^4 \mathbb{1}\left\{q(G_n, t) > M^{1/3}\right\} dt\right]^{3/4} \\
&+ \left[\int_0^1 q(F_n, t)^4 \mathbb{1}\left\{q(F_n, t) > M\right\} dt \int_0^1 q(G_n, t)^4 dt\right]^{1/4}.
\end{aligned}$$

We conclude as above.