WORKING PAPERS

N° 1602

December 2024

“Linear Regressions with Combined Data”

Xavier D'Haultfoeuille, Christophe Gaillac and Arnaud Maurel

Toulouse
School of
Economics

Linear Regressions with Combined Data*

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

December 15, 2024

Abstract

We study best linear predictions 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 and we derive a constructive characterization of the sharp identified set. We then build on this characterization to develop computationally simple and asymptotically normal estimators of the corresponding bounds. We show that these estimators exhibit good finite sample performances.

Keywords: Best linear prediction; data combination; partial identification; inference.

[*] We thank Stephen Hansen, Marc Henry, Toru Kitagawa, Matt Masten, David Pacini and participants at seminars and conferences at Penn State University, TSE, University of Geneva, the Encounters in Econometric Theory 2024, the Munich Econometrics Workshop 2024, 2024 ESEM (Rotterdam), and the Aarhus Workshop in Econometrics V. Arnaud Maurel thanks funding from the Agence Nationale de la Recherche under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

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

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

[§] Duke University and TSE, NBER and IZA, arnaud.maurel@duke.edu.

1 Introduction

Regressions run in applied economics are often not the ideal regressions researchers would like to consider. Notably, it is often the case that the outcome Y and covariates X of interest do not appear in the same dataset. A leading example is intergenerational studies (e.g., intergenerational income or wealth mobility), in common situations where one cannot link parents' and children's outcomes. Another important example is the measurement of racial inequality. For instance, in the context of innovation and patent approval, the dataset of patent applications typically does not include the race of applicants, making it impossible to directly measure racial inequality. A third example is randomized experiments, where the effect of the treatment is measured in the short run but not in the long run, while other databases measure such long-run outcomes. Besides, even if the outcome and covariates of interest do appear in the same dataset, key control variables may be missing from the data. For instance, when measuring the wage returns to education, one may wish to control for a measure of cognitive skills, but this measure may not be available in the main labor market database, even though it appears in another dataset.

When confronted with such data issues, researchers have traditionally relied on imputation methods. Even though this practice is both simple and intuitive, it is important to recognize that it implicitly relies on an exclusion restriction. Such exclusion restrictions may often be questionable. In this paper, we study identification and inference on the regression coefficients in a data combination environment, when relaxing such exclusion restrictions. We consider a general set-up where 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 other variables, that researchers do not seek to include in the regression but that appear in both datasets. For instance, if a common variable is a proxy for X_o, it may be more natural to focus on an "ideal" regression of Y on X_o, but without controlling for that common variable. We denote the set of common variables, either included or not in the "ideal" regression, by W.

We first consider the special case where $X = X_o$. We show in this case that the identified set is nonempty, convex, compact, and derive a simple expression for its support function. The support function expression generalizes the well-known Cambanis-Simons-Stout inequality to a multidimensional case with multiple covariates. When common variables W are also present, one can apply Frisch-Waugh to this setup and partial out Y and X_o. We thus easily extend the previous identification results to account for such common variables. A novel insight in this context is that, in the presence of common variables that are not common regressors, identification gains can be large, in particular making it possible to identify the signs of the coefficients of interest.

We then focus on situations where one is interested in specific components of the parameter vector (or a linear combination thereof), and propose to estimate the identified sets using simple and computationally tractable plug-in estimators. We establish asymptotic normality of the estimators of the lower and upper bounds of the identified sets by leveraging results from the statistical optimal transport and L-statistics literatures. Simulation results indicate that our inference method exhibits good finite sample properties.

Related literature. 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 recent contributions by Fan et al. (2014), Fan et al. (2016), Buchinsky et al. (2022), Athey et al. (2020), D'Haultfœuille et al. (2024) and Bontemps et al. (2024). Most 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 lies in the fact that we study identification and inference, while these two papers primarily focus on deriving the identification region for the "long regression". Particularly relevant for us in this literature is Pacini (2019), who considers the same problem as us. There are three key differences between his paper and ours. First, he does not consider cases where some of the common variables are not used as common regressors ($W \neq X_c$). These cases are prevalent in practice. Second, and importantly, his bounds turn out not to be sharp when X_o is multivariate, and the differences can be substantial. Third, he does not consider inference. Also related to our work is recent work by Hwang (2022). This paper also considers the case where some regressors are only available in the dataset of Y, a case that we do not consider here. On the other hand, Hwang (2022) maintains the restriction that $W = X_c$. For the data combination environments that are common to Hwang (2022) and our paper, she proposes to use the same bounds as those derived in Pacini (2019).

Our paper is related to, but differs from our previous work (D'Haultfœuille et al., 2024) in several important aspects. In this paper we also consider a similar data combination problem. However, a first important difference is that in our previous work, we did not consider the situation where some of the observed variables are not included in the regression. A second important difference is that in our former paper, 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. Also, from a technical point of view, this restriction on the conditional expectation $E[Y|X]$ implies that we had to rely on entirely different optimal transport results.

Technically speaking, 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. In terms of inference, we establish asymptotic normality of our estimators by linking it to the Wasserstein-2 distance between empirical distributions and relying on statistical properties of this distance, see in particular Fournier and Guillin (2015), Del Barrio et al. (2019) and Berthet et al. (2020). Actually, our result yields a central limit theorem on this distance under conditions that are milder than those of Berthet et al. (2020).

A key focus of our analysis is to develop a tractable estimation and inference method. As such, our paper fits into a literature that derives tractable computational methods for partially identified models (see Molinari, 2020 for a recent survey). In particular, we add to the literature that uses tools from optimal transport to devise computationally tractable identification and inference methods for partially identified models (Galichon and Henry, 2011; Galichon, 2018; D'Haultfœuille et al., 2024).

Finally, our paper also speaks to a large and growing empirical literature that deals with similar data combination problems to the one considered in this paper. In particular, the literature on intergenerational income mobility often faces the unavailability of linked income data across generations and relies on exclusion restrictions (see Santavirta and Stuhler, 2022, for a recent survey). Data combination issues are also pervasive in consumption research, as income and consumption are often measured in two different datasets (Crossley et al., 2022). Similar data combination problems frequently arise in the economics of education and returns to skill estimation (Piatek and Pinger, 2016; Garcia et al., 2020; Hanushek et al., 2021), health (Manski, 2018) and labor (Athey et al., 2020). These issues frequently arise also in the context of racial gap in science and innovation (see, e.g., Kerr, 2008; Dossi, 2023; Antman et al., 2024). Since many datasets do not record race together with outcomes of interest such as successful patent applications, race is typically imputed using commonly observed demographics such as last names (Dossi, 2023). The methods we devise in this paper are broadly applicable in these different contexts, allowing researchers to relax the exclusion restrictions that are typically maintained to achieve point-identification of the parameters of interest.

Outline. Section 2 presents the set-up, including our maintained assumptions, and our identification results. Numerical illustrations emphasize that the sharp bounds can be tight in practice. Section 3 develops estimators of the sharp bounds, and inference on the true parameters of interest. This section also derives asymptotic normality of the estimators and construct confidence intervals based on this asymptotic normality. Section 4 examines the finite sample properties of our estimators and confidence intervals through Monte Carlo simulations. Section 5 concludes. The appendix includes a discussion of the sharpness of the bounds of Pacini (2019) and gathers all the proofs of our results.

2 Identification

2.1 Set-up and notation

We seek to identify the best linear predictor $EL(Y|X)$ of Y by $X \in \mathbb{R}^p$, with $X = (X_o', X_c')'$ and $X_k \in \mathbb{R}^{p_k}$ for $k \in \{o, c\}$. 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 $W \in \mathbb{R}^q$ is a vector including X_c. We call X_o the "outside" regressors and X_c the "common regressors". W may include other components than X_c, to capture 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. We thus refer to W as "common variables". Importantly, variables in W but not in X_c 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 $\max(E(Y^2), E(\|X_o\|^2), E(\|W\|^2)) < \infty$ *and* $E(XX')$ *and* $E(WW')$ *are nonsingular.*

Let $b_0 \in \mathbb{R}^p$ be such that $EL(Y|X) = X'b_0$. We seek to characterize the identified set of b_0, defined by

$$\mathcal{B} = \left\{ b \in \mathbb{R}^p : \exists F_{\widetilde{W},\widetilde{X}_o,\widetilde{Y}} :\ F_{\widetilde{W},\widetilde{X}_o} = F_{W,X_o},\ F_{\widetilde{W},\widetilde{Y}} = F_{W,Y},\ EL(\widetilde{Y}|\widetilde{X}) = \widetilde{X}'b \right\}.$$

In words, $\mathcal{B}$ is the set of coefficients of a "long" regression of $\widetilde{Y}$ on $\widetilde{X}$, where the joint distribution of $\widetilde{Y}$ and $\widetilde{X}$ is compatible with the observed marginal distributions. Oftentimes, researchers are interested in specific components of b_0, rather than the whole vector b_0. Therefore, in the following we seek to characterize the corresponding identified set $\mathcal{B}_d = \{d'b :\ b \in \mathcal{B}\}$, for any $d \in \mathbb{R}^p$, and its corresponding bounds:

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

Consider for instance $b_{1,0}$, the first component of b_0. Then, if we let $d = [1, 0, ..., 0]'$, $\mathcal{B}_d$ is the identified set of $b_{1,0}$ and $\overline{b}_d$ and $\underline{b}_d$ are its sharp (upper and lower) bounds. We focus in the following solely on $\overline{b}_d$, which is without loss of generality since $\underline{b}_d = -\overline{b}_{-d}$.

In what follows, we first consider the simplest case with no common variables. Then, we show how common variables affect identification.[1] There are other data combination cases that we do not consider here. One possibility, considered by Hwang (2022),

[1] One could argue that we always have common variables, since we can always let $X_c = 1$. The case without common variable we consider below actually corresponds to $X_c = 1$, whereas in cases with common variable, W is not reduced to $W = 1$ (but W is always assumed to include 1).

is that one observes (Y, X_i, X_c) in one dataset and (X_o, X_c) in another (in such a setup, one could also replace X_c by W). Another possibility, considered by Kitawaga and Sawada (2023), is that one observes (Y, X_1, X_c) in one dataset and (Y, X_2, X_c) in another. Still another possibility is to observe (Y, X_c), (X_1, X_c) and (X_2, X_c) separately. This latter case is partly considered by Moon (2024) when X_1, X_2, X_c are discrete with finite support. It extends Cross and Manski (2002) allowing for more general setups of aggregate data across groups X_c. As we shall see, the setup we consider, in addition to being very common in practice, has the advantage of leading to very simple bounds on b_d.

Finally, we introduce some notation here. For any convex set $C \subseteq \mathbb{R}^p$, we denote its support function by σ_C:

$$\sigma_C(d) := \sup_{x \in C} x'd \quad \forall\, d \in \mathbb{R}^p.$$

We recall that σ_C characterizes C. Also, for any random variables A and B, we let F_A denote the cumulative distribution function (cdf) of A and $F_{A|B}$ denote 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 Supp(A) (resp. 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 obtained by removing v_k from v. We also let $e_{k,r}$ denote the k-th canonical vector of $\mathbb{R}^r$. Finally, for any set S, we let $|S|$ denote its cardinal.

2.2 No common variables

We first assume that there is no common variable ($W = X_c = 1$), so that $X = (X_o', 1)'$. Our main result shows that $\mathcal{B}$ is convex and compact, and characterizes $\overline{b}_d$ for any $d \in \mathbb{R}^p$, $d \neq 0$. Below, we introduce the variable η_d as follows. First, let $(d_2, ..., d_p)$ be $(p-1)$ vectors in $\mathbb{R}^{p-1}$ 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 $p_o = 1$ 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 nonempty, convex, and 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 $d \in \mathbb{R}^p$, $d \neq 0$, $\mathcal{B}_d = [\underline{b}_d, \overline{b}_d]$, with

$$\begin{aligned}\overline{b}_d =& E\left[F^{-1}_{d'E(XX')^{-1}X}(U)F^{-1}_Y(U)\right] \\ =& \frac{E[F^{-1}_{\eta_d}(U)F^{-1}_Y(U)]}{E(\eta_d^2)}.\end{aligned} \quad (1)$$

The first part of the theorem states that $\mathcal{B}$ is a convex, compact set included in the ellipsoid $\mathcal{E}$. In particular, we always have $(0, ..., 0, E[Y|])' \in \mathcal{B}$. This could be expected since, in the absence of common variables, we can always rationalize that Y and X are independent. Also, remark that since the identified set $\mathcal{B}$ is convex and $\overline{b}_d = \sigma_{\mathcal{B}}(d)$, the knowledge of $\overline{b}_d$ for all $d \in \mathbb{R}^p$ ($d \neq 0$) characterizes $\mathcal{B}$.

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

$$\overline{b}_d = \frac{E\left[(F^{-1}_{X_o}(U) - E(X_o))F^{-1}_Y(U)\right]}{V(X_o)}. \quad (2)$$

On the other hand, the true coefficient satisfies $b_d = b_0 = E[(X_o - E(X_o))Y]/V(X_o)$. Thus, (2) indicates that the sharp upper bound on the unknown term $E[X_oY]$ is $E[F^{-1}_{X_o}(U)F^{-1}_Y(U)]$. This is well-known, and corrresponds to the so-called Cambanis-Simons-Stout inequality (see Cambanis et al., 1976). The logic is that (i) $F^{-1}_{X_o}(U)$ and $F^{-1}_Y(U)$ 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 $F^{-1}_{X_o}(U)$ and $F^{-1}_Y(U)$ corresponds to the Fréchet-Hoeffding upper bound.

With multiple regressors, (1) cannot be directly deduced from the Cambanis-Simons-Stout inequality. To get some intuition on (1), 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 Frisch-Waugh's theorem. Now, if we only know the marginal distributions of η_d and Y, the numerator in (1) is simply the upper bound of $E[\eta_d Y]$. That the sharp upper bound $\overline{b}_d$ satisfies (1) 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 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 because basically, the conditional distribution of X_{-1} does not carry any additional information about $E[\eta_d Y]$. Though this can be deduced from Lemma 3.3 in Delon et al. (2023), our own proof is constructive.

Note that Pacini (2019) also obtains an expression for $\sigma_{\mathcal{B}}(d)$, see his Theorem 1.[2] However, when X is multivariate, it turns out that this expression is only an upper bound on the true support function. 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 be substantially tighter than the bounds given in Pacini (2019).

2.3 Common variables

2.3.1 Main result

We now turn to the situation where some covariates are observed in both datasets. In this context, we 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 and for any $d \in \mathbb{R}^p$,*

$$\sigma_{\mathcal{B}}(d) = \frac{1}{E(\eta_d^2)} \left\{\delta_d' E(WW')\delta_Y + E\left[F_{\nu_d|W}^{-1}(U|W) F_{\nu_Y|W}^{-1}(U|W)\right]\right\}. \tag{3}$$

Moreover, for any function g,

$$\sigma_{\mathcal{B}}(d) \le \frac{1}{E(\eta_d^2)} \left\{\delta_d' E(WW')\delta_Y + E[F_{\nu_d|g(W)}^{-1}(U|g(W)) F_{\nu_Y|g(W)}^{-1}(U|g(W))]\right\}, \tag{4}$$

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

[2] See also Proposition 2 of Hwang (2022) for bounds on each component of b_0. These bounds turn out to be the same as those of Pacini (2019).

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 $\sigma_{\mathcal{B}}(d)$ 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 discrete support, the outer bound reported in Equation 4 is elementary to estimate and does not suffer from any curse of dimensionality. Moreover, this bound turns out to be sharp when $\nu_Y \perp\!\!\!\perp W|g(W)$ and $\nu_d \perp\!\!\!\perp W|g(W)$.

Not surprisingly, the bounds depend on which components of W are in X_c. To understand this, assume first that $W = X_c$, namely that we do not have any additional variables excluded from the linear regression of interest. Then, as in the case without regressors, 0 always belongs to the identified set for X_o. To see this, let $\mathcal{B}_o$ denote the identified set of b_o and let $\eta_o = X_o - EL[X_o|X_c]$. Then, by Frisch-Waugh theorem, $b_o = V(\eta_o)^{-1}E[\eta_o \nu_Y]$. As a result, by applying Theorem 1 to (η_o, ν_Y) conditional on W and integrating over W, we obtain that $\mathcal{B}_o$ is convex with support function

$$\sigma_{\mathcal{B}_o}(d) = E[F^{-1}_{d'V(\eta_o)^{-1}\eta_o|X_c}(U|X_c)F^{-1}_{\nu_Y|X_c}(U|X_c)].$$

The right-hand side is non-negative for all d, which shows that $0 \in \mathcal{B}_o$.[3]

On the other hand, if no components of W are included in the linear regression of interest (no X_c), $\mathcal{B}_o$ may not include 0, since it may be that for some $d \in \mathbb{R}^p$,

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

For instance, if $(W, X_o) \sim \mathcal{N}(0, \Sigma)$ and $(W, Y) \sim \mathcal{N}(0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{5}$$

then $\sigma_{\mathcal{C}}(-1) = 1 - 2\rho^2 < 0$ if $\rho > 1/\sqrt{2}$.

[3] On the other hand, 0 may not be in the identified set of b_c, the 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.

2.3.2 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. More generally, we could assume that we only observe the distributions of $(W, Y)|D = 1$ and $(W, X_o)|D = 0$ for some binary variable D. In this framework, we have considered up to now 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 the corresponding test is rejected, we can weaken the independence assumption by assuming instead that

$$(X_o, Y) \perp\!\!\!\perp D|W, \quad p := P(D = 1) \text{ is known.} \tag{6}$$

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 of (6) 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 (6) 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 η_Y), we cannot directly regress Y on W conditional on $D = 1$. However, we can obtain it by considering a weighted regression, since

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

We can obtain δ_d (and then η_d) similarly, using the weights $(1 - D)/(1 - P(W))$. Then, Equation (3) is replaced by:

$$\sigma_{\mathcal{B}}(d) = \frac{1}{E\left[\frac{(1-D)\eta_d^2}{1-P(W)}\right]} \left\{\delta_d' E(WW')\delta_Y + E\left[F^{-1}_{\nu_d|W,D=0}(U|W)F^{-1}_{\nu_Y|W,D=1}(U|W)\right]\right\}.$$

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 $EL(Y|X)$, $EL(Y|X, D = 1)$. If so, δ_Y is now $E[WW'|D = 1]^{-1}E[WY|D =$

1] and is thus obtained by an unweighted regression, whereas δ_d (and then η_d) is obtained by regressing η_d on W with weights $P(W)/(1-P(W))$. The upper bound $\overline{b}_d$ becomes

$$\overline{b}_d = \frac{E(D)}{E\left[\frac{(1-D)P(W)\eta_d^2}{1-P(W)}\right]}\Bigg\{\delta_d' E(WW'|D=1)\delta_Y \\ + E\left[F^{-1}_{\nu_d|W,D=0}(U|W)F^{-1}_{\nu_Y|W,D=1}(U|W)|D=1\right]\Bigg\}. \qquad (7)$$

Finally, in some cases one of the 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). If so and we focus as above on $EL(Y|X, D = 1)$, we obtain a similar upper bound on $\overline{b}_d$ as in (7), with just a few differences. First, δ_d (and then η_d) is obtained by regressing η_d on W with weights $P(W)$. Second, we now have

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

Note that in this case and the one before, we do not require the joint independence condition in (6) but only $X_o \perp\!\!\!\perp D|W$.

2.3.3 Additional, non-common variables

In practice, one may have access to additional variables that appear in the dataset of Y only, or in the dataset of X_o only. For instance, let us assume that we identify the distributions of (W, Y, Z) on the one hand and (W, X_o) on the other hand. The identified set of b_0 then becomes

$$\mathcal{B}_Z = \Bigg\{b \in \mathbb{R}^p : \exists F_{\widetilde{W},\widetilde{X}_o,\widetilde{Y},\widetilde{Z}} : \ F_{\widetilde{W},\widetilde{X}_o} = F_{W,X_o}, \ F_{\widetilde{W},\widetilde{Y},\widetilde{Z}} = F_{W,Y,Z}, EL(\widetilde{Y}|\widetilde{X}) = \widetilde{X}'b\Bigg\}.$$

The following proposition shows that the knowledge of the conditional distribution of $Z|W, Y$ is actually useless in terms of identification.

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

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

2.4 Linear inequality restrictions

Oftentimes, a priori information on b_0 is available. For instance, theory could imply that some components of b_0 are nonpositive or nonnegative. We study here the identified set of b_d under the additional constraints that $Rb_0 \leq r$, where R is a $r \times p$ matrix, r is a column vector of size r and "$\leq$" should be understood componentwise. Then, the identified set of b_0 satisfies

$$\mathcal{B}^c := \mathcal{B} \cap \{b : Rb \leq r\},$$

where $\mathcal{B}$ denotes the unconstrained identified set of b_0 obtained as above. As the intersection of a compact, convex set with a closed and convex set, $\mathcal{B}^c$ is compact and convex. Then, $\mathcal{B}^c_d := \{d'b : b \in \mathcal{B}^c\}$ is a compact interval $[\underline{b}^c_d, \overline{b}^c_d]$. Moreover, $b \in \mathcal{B}$ if and only if $u'b \leq \sigma_{\mathcal{B}}(u)$ for all $u \in \mathbb{S}$, with $\mathbb{S}$ the unit sphere of $\mathbb{R}^p$. Hence,[4]

$$\overline{b}^c_d = \sup_{b \in \mathbb{R}^p} d'b \quad \text{s.t. } Rb \leq r \text{ and } s'b \leq \sigma_{\mathcal{B}}(s)\ \forall s \in \mathbb{S}. \tag{8}$$

Now, (8) cannot be solved directly because of the infinitely many constraints $s'b \leq \sigma_{\mathcal{B}}(s)$ for all $s \in \mathbb{S}$. Instead, we derive lower and upper bounds on $\overline{b}^c_d$. Note that the motivation for deriving a lower bound on $\overline{b}^c_d$, and not solely an upper bound, is to be able to quantify the quality of the computational approximation of $\overline{b}^c_d$. To construct these bounds, fix $(s_g)_{g=1,...,G} \in \mathbb{S}^G$ and let $S = [s_1, s_2, ..., s_G]'$. Finally, let $\overline{R} := [R'\ S']'$ and let $\overline{r} := [r, \sigma_{\mathcal{B}}(s_1), ..., \sigma_{\mathcal{B}}(s_G)]'$. Then, we obtain the following upper bound $\overline{\overline{b}}^c_d$ on $\overline{b}^c_d$:

$$\overline{\overline{b}}^c_d := \sup_{b \in \mathbb{R}^p} d'b \quad \text{s.t. } \overline{R}b \leq \overline{r}.$$

To obtain a lower bound $\underline{\overline{b}}^c_d$ on $\overline{b}^c_d$, we reason as with inside regressors: (i) construct the convex hull C of $\{(s_g, \sigma_{\mathcal{B}}(s_g))_{g=1,...,G}\}$; (ii) express C as $\underline{R}b \leq \underline{r}$ for some $\underline{R}$, $\underline{r}$; (iii) compute the lower bound

$$\underline{\overline{b}}^c_d := \sup_{b \in \mathbb{R}^p} d'b \quad \text{s.t. } \underline{R}b \leq \underline{r}.$$

[4] Recall that the lower bound satisfies $\underline{b}^c_d = -\overline{b}^c_{-d}$ so we can focus on $\overline{b}^c_d$.

3 Estimation and inference

We now turn to the estimation of $\overline{b}_d$ and inference on b_d, based on i.i.d. samples. We focus hereafter on the case without common variables.

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.*

3.1 Definition and computation of the estimators

Let $\widehat{\eta}_{dj}$ denote j's residual in the (sample) regression of T_1 on T_{-1} (the definition of T is given at the beginning of Section 2.2). 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 the $(Y_i)_{i=1,...,n}$ and the $(\widehat{\eta}_{dj})_{j=1,...,m}$. Recall from Theorem 1 that

$$\overline{b}_d = \frac{\int_0^1 F^{-1}(t)G^{-1}(t)dt}{E(\eta_d^2)}.$$

Then, we consider the following 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 the $(\widehat{\eta}_{dj})_{j=1,...,m}$. Note that we can compute the numerator of $\widehat{\overline{b}}_d$ at a very low cost. To see this, remark that for any real-valued variables U_1, U_2 with finite second moments and cdfs F_1, F_2,

$$\int_0^1 F_1^{-1}(t)F_2^{-1}(t)dt = \frac{1}{2}\left[E[U_1^2] + E[U_2^2] - W_2^2(F_1, F_2)\right], \qquad (9)$$

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 F_n and G_m, we can compute $W_2(F_1, F_2)$ using algorithms to compute the so-called Earth Mover's Distance in one dimension, which have a complexity of $O(m + n)$, see, e.g., Rubner et al. (2000).

3.2 Inference

We now turn to the construction of confidence interval on b_d, based on asymptotic normality. We first establish the asymptotic distribution of $\widehat{\overline{b}}_d$. We focus on the case where both η_d and Y have infinite support; otherwise, the result follows from Del Barrio et al. (2024). We impose Assumption 3 below. Let $\mathcal{D}_f$ denote the points of discontinuity of a function f.

Assumption 3

(i) The distribution of η_d is continuous with respect to the Lebesgue measure, with density g.

(ii) $E(Y^{4+\varepsilon}) < \infty$ and $E(\|X\|^{4+\varepsilon}) < \infty$ for some $\varepsilon > 0$.

(iii) $\mathcal{D}_{F^{-1}} \cap \mathcal{D}_{G^{-1}} = \emptyset$ and there exists $C_1, C_2 > 0$ such that:

$$\frac{g(x)}{G(x)(1-G(x))} \geq C_1 \wedge \frac{C_2}{|x|\ln(1+|x|)^2} \quad \forall\, x \in Supp(\eta_d). \tag{10}$$

The condition $\mathcal{D}_{F^{-1}} \cap \mathcal{D}_{G^{-1}} = \emptyset$ is satisfied if F^{-1} and G^{-1} are continuous, which in turn holds if Supp(Y) and Supp(η_d) are intervals. In particular, F^{-1} may be continuous even if the distributions of Y has mass points. But we can also allow for discontinuities of F^{-1} and G^{-1}, as long as they do not intersect. Condition (10) holds on Supp(η_d) $\cap\, [0, \infty)$ for all distributions that have increasing hazard rates. This includes log-concave distributions, since their survival is then log-concave. It also holds for many distributions with decreasing hazard rates, such as Pareto distributions, whose hazard rate is of the form $1/x$, and Weibull distributions, whose hazard rates is of the kind $\alpha\beta x^{\beta-1}$ with $\alpha, \beta > 0$. More generally, we expect Condition (10) to be mild, since if we denote by $\overline{x}$ the supremum of the support of η_d, we have, for all $A < \overline{x}$ satisfying $G(A) > 0$,

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

On the other hand, gor 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 $g(x)/[G(x)(1-G(x))] \leq C_1 \wedge C_2/(|x|\ln(1+|x|)^2)$ for all x large enough; and similarly one cannot have $g(x)/[G(x)(1-G(x))] \leq C_1 \wedge C_2/(|x|\ln(1+|x|)^2)$ for all x small enough.

To define the asymptotic distribution, we introduce additional objects. First, to simplify notation, let F and G denote respectively the cdf of Y and η_d. Then, let $h = F^{-1} \circ G$ and

$$\begin{aligned}
\psi_1 &:= -\overline{b}_d(\eta_d^2 - E[\eta_d^2]),\\
\psi_2 &:= -E[h(\eta_d)T_{-1}']E[T_{-1}T_{-1}']^{-1}T_{-1}\eta_d,\\
\psi_3 &:= -\int [\mathbb{1}\{\eta_d \leq t\} - G(t)]h(t)dt,\\
\psi_4 &:= -\int [\mathbb{1}\{Y \leq t\} - F(t)]G^{-1} \circ F(t)dt.
\end{aligned}$$

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 $(\eta_{dj})_{j=1,\dots,m}$ (note that G_m cannot be computed in practice, since the $(\eta_{dj})_{j=1,\dots,m}$ are unobserved). Then, let

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

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

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

First, let us 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 such cases, the asymptotic variance V_d becomes simpler. Also, 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 (9) 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.[5] 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

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

tails of the distributions, one can show that their subconditions (FG1) and (FG3) already imply Assumption 3-(i) and (iii) not only for the distribution of η_d but also for that of Y.[6]

Now, let us give a sketch of the proof of Theorem 3. 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

$$\sqrt{m}\int F_n^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt = -E[h(\eta_d)T_{-1}']\sqrt{m}(\widehat{\gamma} - \gamma_0) + o_P(1),$$

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 $(nm/(n+m))^{1/2}\int_0^1(F_n^{-1}(t)G_m^{-1}(t) - F^{-1}(t)G^{-1}(t))dt$. To this end, we use the decomposition

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

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). To show that the remainder term $r_{n,m}$ is negligible, we relate it 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 (10). We do this by linking $W_2(G_m, G)$ with the variance of order statistics, and relying on a lemma similar to Corollary 2.12 in Boucheron and Thomas (2015); see Lemma 2 in Appendix B.5.

Next, we construct our confidence intervals on $\overline{b}_d$ using a plug-in estimator of V_d. Specifically, let $\widehat{g} = F_n^{-1} \circ \widehat{G}_m$ and

$$\widehat{\psi}_{1i} := -\widehat{\overline{b}}_d\left(\widehat{\eta}_{di}^2 - \frac{1}{m}\sum_{j=1}^m \widehat{\eta}_{dj}^2\right),$$

$$\widehat{\psi}_{2i} := -\left(\frac{1}{m}\sum_{j=1}^m \widehat{g}(\widehat{\eta}_{di})T_{-1j}'\right)\left(\frac{1}{m}\sum_{j=1}^m T_{-1j}T_{-1j}'\right)^{-1} T_{-i}\widehat{\eta}_{di},$$

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

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

Then, define

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

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

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

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 $z_{1-\alpha}$ here since under the conditions above, the identified interval of b_d is not reduced to a singleton: $\overline{b}_d > 0 > -\overline{b}_{-d}$.

4 Simulations

To illustrate the finite sample properties of our estimator, we consider the following DGPs:

- DGP1: $Y = X_o'b_0 + \varepsilon$, with $X_o = (X_{o,1}, X_{o,2}) = (\exp(N_1), N_2)$, $N = (N_1, N_2) \sim \mathcal{N}(0, \Sigma)$, Σ as in (5) with $\rho = 0.3$, $\varepsilon \sim \mathcal{N}(0,1)$ and $b_0 = (1,1)$.
- DGP2: $Y = X_{o,1}b_{0,1} + X_{o,2}b_{0,2} + X_{o,1}X_{o,2}b_{0,3} + X_{o,1}^2 b_{0,4} + X_{o,2}^2 b_{0,5} + \varepsilon$, with $(X_{o,1}, X_{o,2}) = N$ as in DGP1 and $b_0 = (1, 1, 2, 0.5, -0.8)$.
- DGP3: Same as DGP2, except that $b_0 = (1, 1, 2, 0, 0)$.

We assume that the two samples have the same size ($n = m$), which varies between 400 and 4,800. In Table 1 we report the estimated identified set and the average bounds, across 500 simulations, of the 95% confidence intervals for $b_{0,1}$ with each of the different sample sizes. The first columns report results obtained with our confidence interval $\text{CI}_{1-\alpha}$ defined above. We also use the standard bootstrap as an alternative. We report what we call the excess length ("Excess Length"), namely the mean difference between the length of the confidence sets and that of the identified set. We also report the coverage rates across simulations ("Coverage"). This 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.

	Asymptotic normality			Bootstrap		
	Bounds	EL	Coverage	Bounds	EL	Coverage
DGP1						
Identified	[-0.809,1.275]					
400	[-1.146,1.683]	0.745	0.912	[-1.076,1.628]	0.62	0.842
800	[-1.073,1.601]	0.590	0.914	[-1.011,1.548]	0.475	0.864
1600	[-1.008,1.526]	0.450	0.924	[-0.973,1.492]	0.381	0.868
2400	[-0.969,1.492]	0.378	0.918	[-0.967,1.491]	0.374	0.930
4800	[-0.934,1.435]	0.285	0.934	[-0.923,1.433]	0.273	0.928
DGP2						
Identified	[-3.007,3.007]					
400	[-3.327,3.325]	0.637	0.900	[-3.317,3.318]	0.621	0.916
800	[-3.236,3.235]	0.456	0.932	[-3.234,3.234]	0.454	0.916
1600	[-3.166,3.166]	0.317	0.940	[-3.167,3.167]	0.319	0.914
2400	[-3.151,3.152]	0.289	0.948	[-3.146,3.146]	0.277	0.950
4800	[-3.105,3.105]	0.195	0.956	[-3.102,3.101]	0.188	0.950
DGP3						
Identified	[-2.802,2.802]					
400	[-3.106,3.104]	0.606	0.900	[-3.094,3.095]	0.586	0.896
800	[-3.018,3.016]	0.430	0.926	[-3.021,3.022]	0.439	0.916
1600	[-2.954,2.954]	0.304	0.942	[-2.952,2.954]	0.303	0.908
2400	[-2.94,2.941]	0.277	0.938	[-2.934,2.934]	0.264	0.928
4800	[-2.896,2.896]	0.189	0.944	[-2.892,2.892]	0.181	0.946

Notes: results obtained with 500 simulations. 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 bounds of the 95% confidence intervals over simulations. "EL" is the excess length, i.e. the average length of the confidence region minus the length of the identified set. Column "Coverage" displays the minimum, over $b = (b_1, ..., b_p) \in \mathcal{B}$, of the estimated probability that $b_1 \in \text{CR}_{1-\alpha}(b_{0,1})$. We use 1,000 bootstrap samples to compute the confidence intervals.

Table 1: Monte Carlo simulations 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 and approximately at the $n^{-1/2}$ rate for the three DGPs we consider. Second, the confidence intervals based on asymptotic normality and the bootstrap are similar in terms of length and coverage, with coverage rates converging to 95% for the three DGPs. If anything, intervals based on asymptotic normality seem to produce coverage rates slightly closer to the nominal rate of 95%. This is especially the case for DGP1, for which distortions in coverage rates are the largest.

5 Conclusion

We study best linear predictions 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 fields in empirical economics. A common approach has been to rely on imputation methods, which rely on exclusion restrictions. We take another route and derive a constructive characterization of the sharp identified set. We use this characterization to build asymptotically normal estimators of the corresponding bounds. Monte Carlo simulation exercises indicate that our estimators, which can be computed at a very limited computational cost, exhibit good finite sample performances.

References

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.

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.

Bontemps, C., J.-P. Florens, and N. Meddahi (2024). Functional ecological inference. *Journal of Econometrics forthcoming.*

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.

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.

Dossi, G. (2023). Race and science. Working Paper.

D'Haultfœuille, X., C. Gaillac, and A. Maurel (2024). Partially linear models under data combination. *Review of Economic Studies forthcoming*.

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

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.

Galichon, A. (2018). *Optimal transport methods in economics.* Princeton University Press.

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

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

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

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

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.

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

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

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

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.

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 (2022). Name-based estimators of intergenerational mobility. Mimeo.

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

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

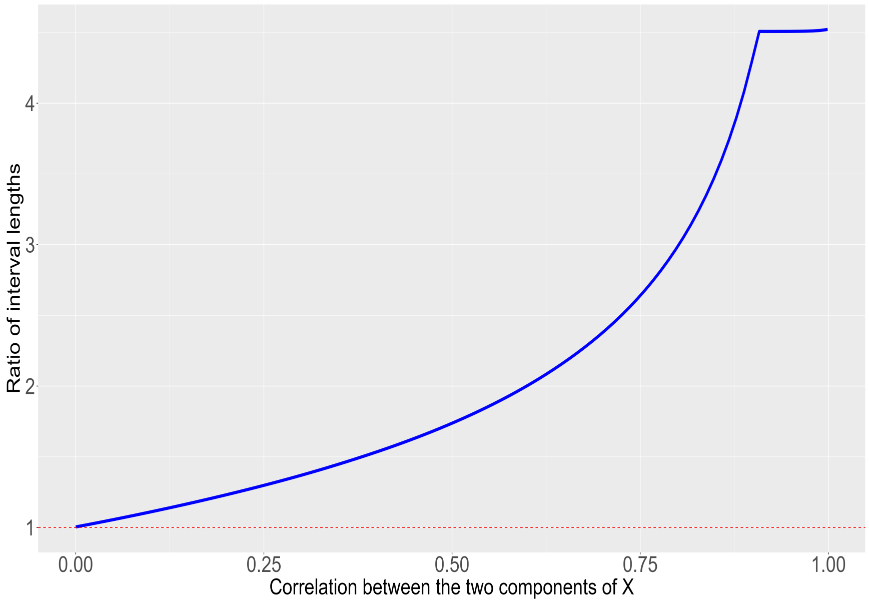
Pacini (2019) gives the expression of the support function of the identified set of b_0 in the case without inside regressors (but allowing for common regressors, denoted by z in his paper). His bounds coincide with ours when X_o is univariate, but not otherwise. In the multidimensional case, his expression of $\sigma_{\mathcal{B}}$ is an upper bound of the true support function. 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 tha 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 in (5) (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 1 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 1: Comparison between Pacini (2019) bounds and the sharp bounds

B Proofs

B.1 Theorem 1

First, if $b \in \mathcal{B}$, then $EL(\widetilde{Y}|\widetilde{X}) = \widetilde{X}'b$. 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,

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

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

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

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 $D \sim$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

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

$$=pE\left[\widetilde{X}_1\widetilde{Y}_1\right]+(1-p)E\left[\widetilde{X}_2\widetilde{Y}_2\right]$$
$$=E(XX')(pb_1+(1-p)b_2).$$

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

Now, we prove $\sigma_{\mathcal{B}}(d)=E[F^{-1}_{d'E[XX']^{-1}X}(U)F^{-1}_Y(U)]$. We have

$$\sigma_{\mathcal{B}}(d)=\max_{\Pi\in\mathcal{M}(F_X,F_Y)}\int\left[d'E[XX']^{-1}x\right]\,y\,d\Pi(x,y), \tag{11}$$

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

$$\sigma_{\mathcal{B}}(d)\leq\max_{\Pi\in\mathcal{M}(F_{X_d},F_Y)}\int uyd\Pi(u,y).$$

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

$$\max_{\Pi\in\mathcal{M}(F_{X_d},F_Y)}\int uyd\Pi(u,y)=E[F^{-1}_{X_d}(U)F^{-1}_Y(U)]. \tag{12}$$

Hence, $\sigma_{\mathcal{B}}(d)\leq E[F^{-1}_{X_d}(U)F^{-1}_Y(U)]$.

Now, for any $U\sim\mathcal{U}([0,1])$, let $\widetilde{Y}=F^{-1}_Y(U)$. Let also C denote a copula of $M'E[XX']^{-1}X$ (recall the construction of M at the beginning of Section 2.2) 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 $\sigma_{\mathcal{B}}(d)$, $\sigma_{\mathcal{B}}(d)\geq E[F^{-1}_{X_d}(U)F^{-1}_Y(U)]$. Equation (1) follows.

Finally, we prove (1). 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,

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

B.2 Theorem 2

By construction, $EL(Y|X) = E[X_d Y]$. The exact same reasoning as in the proof of Theorem 1 shows that the identified set $\mathcal{B}(w)$ of $E[X_d Y | W = w]$ is convex. By integrating over w, $\mathcal{B}$ is thus convex. Let U be such that $U|W$ is uniform. Then, the support function of $\mathcal{B}(w)$ satisfies

$$\begin{aligned}\sigma_{\mathcal{B}(w)}(d) =& E\left[F^{-1}_{W'\delta_d + \nu_d | W}(U|W) F^{-1}_{W'\delta_Y + \nu_Y | W}(U|W) | W = w\right] \\ =& E\left[\left(W'\delta_d + F^{-1}_{\nu_d|W}(U|W)\right)\left(W'\delta_Y + F^{-1}_{\nu_Y|W}(U|W)\right) | W = w\right].\end{aligned}$$

Next, $E[X_d Y] = E[E[X_d Y | W]] \leq E[\sigma_{\mathcal{B}(W)}(d)]$. Moreover, the bound is reached by considering $(X_d, Y) = (F^{-1}_{W'\delta_d + \nu_d|W}(U|W), F^{-1}_{W'\delta_Y + \nu_Y|W}(U|W))$. Thus, $\sigma_{\mathcal{B}}(d) = E[\sigma_{\mathcal{B}(W)}(d)]$. Then, since $(W, F^{-1}_{\nu_d|W}(U|W))$ has the same distribution as (W, ν_d); and similarly with ν_Y instead of ν_d,

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

The first point of the proposition follows.

To obtain the second point, remark that

$$\begin{aligned}E[X_d Y] =& E\left[(W'\delta_d + \nu_d)(W'\delta_Y + \nu_Y)\right] \\ =& \delta_d E\left[WW'\right] \delta_Y + E[\nu_d \nu_Y]\end{aligned}$$

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

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 $\sigma_{\mathcal{B}}(d)$. The third point of the proposition follows.

B.3 Proposition 1

Let us denote by $\mathcal{B}_Z(w)$ the identified set of $E[X_d Y|W = w]$ when observing Z, whereas $\mathcal{B}(w)$ still denotes the identified set of $E[X_d Y|W = w]$ 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 $E[X_d Y|W = w]$ is convex. Thus, it is characterized by its support function $\sigma_{\mathcal{B}_Z(w)}$. As in (11), 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 $w = (x_c, w_1)$. By Lemma 3.3 of Delon et al. (2023),

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

Hence, by integrating over w, we obtain $\sigma_{\mathcal{B}_Z} = \sigma_{\mathcal{B}}$. The result follows.

B.4 Theorem 3

Linear approximation of the first terms

We first show that

$$
\begin{aligned}
\sqrt{\frac{nm}{n+m}}\left(\widehat{\overline{b}}_d - \overline{b}_d\right) =& \frac{1}{E(\eta_d^2)}\left[\sqrt{\frac{nm}{n+m}}\int_0^1 (F_n^{-1}G_m^{-1} - F^{-1}G^{-1})dt + \frac{\sqrt{\lambda}}{m^{1/2}}\sum_{i=1}^m \psi_{1i} + \psi_{2i}\right] \\
&+ o_P(1). \qquad (13)
\end{aligned}
$$

First, remark that

$$\widehat{\overline{b}}_d - \overline{b}_d = \frac{1}{\widehat{E}(\widehat{\eta}_d^2)}\left[\int_0^1 (F_n^{-1}\widehat{G}_m^{-1} - F^{-1}G^{-1})dt - \overline{b}_d(\widehat{E}(\widehat{\eta}_d^2) - E(\eta_d^2))\right]. \qquad (14)$$

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

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

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 (14), $n/(n+m) \to \lambda$ and the definition of ψ_1, this yields

$$\begin{aligned}\sqrt{\frac{nm}{n+m}}\left(\widehat{\overline{b}}_d - \overline{b}_d\right) = & \frac{1}{E(\eta_d^2)}\left[\sqrt{\frac{nm}{n+m}}\int_0^1 (F_n^{-1}\widehat{G}_m^{-1} - F^{-1}G^{-1})dt + \frac{\sqrt{\lambda}}{m^{1/2}}\sum_{i=1}^m \psi_{1i}\right] \\ & + o_P(1). \end{aligned} \qquad (15)$$

Let us now prove that

$$\sqrt{m}\int_0^1 F_n^{-1}(\widehat{G}_m^{-1} - G_m^{-1})dt = -E[h(\eta_d)T'_{-1}]\sqrt{m}(\widehat{\gamma} - \gamma_0) + o_P(1). \qquad (16)$$

When combined with (15), 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 (13).

Let σ_1 (resp. σ_2) denote a permutation of $\{1, ..., m\}$ such that $\eta_{d\sigma_1(1)} \leq ... \leq \eta_{d\sigma_1(m)}$ (resp. $\widehat{\eta}_{d\sigma_1(1)} \leq ... \leq \widehat{\eta}_{d\sigma_1(m)}$) and 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,

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

Next, remark that

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

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

$$\begin{aligned} -\left[\int_0^1 F_n^{-1}Q_{1m}'dt\right](\widehat{\gamma}-\gamma_0) \leq & \int_0^1 F_n^{-1}(\widehat{G}_m^{-1}-G_m^{-1})dt \\ & \leq -\left[\int_0^1 F_n^{-1}Q_{2m}'dt\right](\widehat{\gamma}-\gamma_0). \end{aligned} \quad (17)$$

Now, let $\tilde{Y}_i = h(\eta_{di})$. Because G is continuous, $\tilde{Y}_i$ has cdf F. Let $\tilde{F}_m$ denote the empirical cdf of $(\tilde{Y}_i)_{i=1,\dots,m}$. Let $W_2(F,G) := \left(\int_0^1 [F^{-1}(t)-G^{-1}(t)]^2 dt\right)^{1/2}$. Then,

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

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). Hence, we have, for $k \in \{1,2\}$

$$\begin{aligned} \left\|\int_0^1 \left(F_n^{-1}-\tilde{F}_m^{-1}\right)Q_{km}'dt\right\| \leq & \left(\int_0^1 \left[F_n^{-1}-\tilde{F}_m^{-1}\right]^2 dt\right)^{1/2}\left(\int_0^1 \|Q_{km}\|^2 dt\right)^{1/2} \\ & = o_P(1). \end{aligned} \quad (18)$$

Next, remark that

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

Together with (18), this proves that

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

Using (18) again but with $j=2$ and (17), (16) follows provided that

$$\int_0^1 \tilde{F}_m^{-1}Q_{2m}dt \xrightarrow{P} E[h(\eta_{di})T_{-1i}]. \quad (19)$$

Fix $\delta > 0$. Let

$$U_{mi} := \mathbb{1}\left\{|h(\eta_{di}) - h(\widehat{\eta}_{di})| < \frac{\delta}{10E[\|T_{-1}\|^2]^{1/2}}\right\}.$$

We have

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

Consider T_0. By Cauchy-Schwarz inequality,

$$\|T_0\| \leq \left(\frac{1}{m}\sum_{i=1}^m h(\eta_{di})^2\right)^{1/2} \left(\frac{1}{m}\sum_{i=1}^m \|T_{-1i}\|^2 (1 - U_{mi})\right)^{1/2}. \tag{20}$$

By the dominated convergence theorem, there exists $M > 0$ such that

$$E[h(\eta_d)^2] E\left[\|T_{-1}\|^2 \mathbb{1}\left\{|\eta| > M\right\}\right] < \frac{\delta^2}{16}. \tag{21}$$

Moreover, since g is continuous, there exists $c > 0$ such that if $|\eta_{di}| \leq M$ and $|\eta_{di} - \widehat{\eta}_{di}| < c$, then $U_{mi} = 1$. As a result,

$$1 - U_{mi} \leq \mathbb{1}\left\{|\eta_{di}| > M\right\} + (1 - I_{m,c}), \tag{22}$$

with $I_{m,c} := \mathbb{1}\left\{\max_{i=1,\dots,m} |\widehat{\eta}_{di} - \eta_{di}| < c\right\}$. Besides,

$$\begin{aligned}\max_{i=1,\dots,m} |\widehat{\eta}_{di} - \eta_{di}| &= \max_{i=1,\dots,m} \left|T'_{-1i}(\widehat{\gamma} - \gamma_0)\right| \\ &\leq \left[\max_{i=1,\dots,m} \|T_{-1i}\|\right] \|\widehat{\gamma} - \gamma_0\| \\ &= o_P(n^{1/2}) \times O_P(n^{-1/2}) \\ &= o_P(1). \end{aligned} \tag{23}$$

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). By combining the law of large numbers with (20)-(23), we obtain, with probability approaching one (wpao),

$$\|T_0\| \leq \frac{\delta}{3}. \tag{24}$$

Next, consider T_1. We have

$$\begin{aligned} T_1 &= \frac{1}{m}\sum_{i=1}^{m} h(\widehat{\eta}_{di}) T_{-1i} U_{mi} \\ &= \frac{1}{m}\sum_{i=1}^{m} \left[h(\widehat{\eta}_{di}) - h(\eta_{di})\right] T_{-1i} U_{mi} + \frac{1}{m}\sum_{i=1}^{m} h(\eta_{di}) T_{-1i} - \frac{1}{m}\sum_{i=1}^{m} h(\eta_{di}) T_{-1i}(1 - U_{mi}) \\ &=: T_{11} + T_{12} - T_{13}. \end{aligned}$$

By the law of large numbers,

$$T_{12} \xrightarrow{P} E[h(\eta_d) T_{-1}]. \tag{25}$$

By Cauchy-Schwarz inequality, we obtain for T_{13} the same inequality as (20). Thus, wpao,

$$\|T_{13}\| \leq \frac{\delta}{9}. \tag{26}$$

Turning to T_{11}. we have

$$\begin{aligned} T_{11} &\leq \left(\frac{1}{m}\sum_{i=1}^{m} [h(\widehat{\eta}_{di}) - h(\eta_{di})]^2 U_{mi}\right)^{1/2} \left(\frac{1}{m}\sum_{i=1}^{m} \|T_{-1i}\|^2\right)^{1/2} \\ &\leq \frac{\delta}{10 E[\|T_{-1}\|^2]^{1/2}} \left(\frac{1}{m}\sum_{i=1}^{m} \|T_{-1i}\|^2\right)^{1/2}, \end{aligned}$$

where the second inequality follows by definition of U_{mi}. Hence, wpao,

$$\|T_{11}\| \leq \frac{\delta}{9}. \tag{27}$$

Thus, by combining the triangle inequality, a union bound and (25)-(27), we obtain that wpao,

$$\|T_1 - E[h(\eta_d) T_{-1}]\| \leq \frac{\delta}{3}. \tag{28}$$

Finally, consider T_2. First,

$$T_2 \leq \left(\frac{1}{m}\sum_{i:U_{m\sigma_2(i)}=1} [h(\widehat{\eta}_{d\sigma_2(i)} - h(\eta_{d\sigma_1(i)})]^2\right)^{1/2} \left(\frac{1}{m}\sum_{i=1}^{m} \|T_{-1i}\|^2\right)^{1/2}.$$

By the rearrangement inequality, because g is increasing,

$$\sum_{i:U_{m\sigma_2(i)}=1} h(\eta_{d\sigma_1(i)}) h(\widehat{\eta}_{d\sigma_2(i)}) \geq \sum_{i:U_{mi}=1} h(\eta_{di}) h(\widehat{\eta}_{di}).$$

Thus,

$$\frac{1}{m}\sum_{i:U_{m\sigma_2(i)}=1}[h(\widehat{\eta}_{d\sigma_2(i)}-h(\eta_{d\sigma_1(i)}))]^2 \leq \frac{1}{m}\sum_{i=1}^{m}[h(\eta_{di})-h(\widehat{\eta}_{di})]^2 U_{mi}$$
$$\leq \frac{\delta^2}{100E[\|T_{-1}\|^2]}.$$

Hence, wpao

$$\|T_2\| \leq \frac{\delta}{3}. \tag{29}$$

Finally, by combining (24), (28) and (29), we obtain that wpao,

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

Equation (16) follows.

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, whereas 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). Now, we apply Theorem 1 in Shorack and Wellner (1986), 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 1-(i), provided that their Assumptions 1 and 2 hold, $T_n - \mu_n$ by their $\tilde{T}_n - \mu$, which is our T_{2n}.

Now, Assumption 3-(ii) implies, by, e.g. Remark 19.1 in Shorack and Wellner (1986), that $|F^{-1}(t)| \leq M_1/[(t(1-t)]^{1/(4+\varepsilon)}$ and $|G^{-1}(t)| \leq M_2[(t(1-t)]^{1/(4+\varepsilon)}$ for some M_1 and M_2. Hence, (16) and (19) in their Assumption 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 $a < 1/2$. Since $J_n = J$

in $\tilde{T}_n$, their Assumption 2 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 $|F^{-1}|$. Becaues G^{-1} is monotone, its set of discontinuities $\mathcal{D}_{G^{-1}}$ is countable. Moreover, by Assumption 3-(iii), we have, for each $x \in \mathcal{D}_{G^{-1}}$, $\mu(\{x\}) = 0$. Hence, their Assumption 2 holds here. Then, by Theorem 1 in Shorack and Wellner (1986) and their equation just above (13),

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

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

$$\sqrt{\frac{nm}{n+m}}T_{2m} = (1-\lambda)^{1/2}\frac{1}{n^{1/2}}\sum_{i=1}^{m}\psi_{4i} + o_P(1). \tag{30}$$

Similarly,

$$\sqrt{\frac{nm}{n+m}}T_{1m} = \lambda^{1/2}\frac{1}{m^{1/2}}\sum_{i=1}^{n}\psi_{3i} + o_P(1). \tag{31}$$

We now show that $R_{n,m} := \sqrt{nm/(n+m)}r_{n,m} = o_P(1)$. Combined with (13), (30) and (31), 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.

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,

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

Theorem 1 in Fournier and Guillin (2015) shows that

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

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

$$E\left[W_2^2(G_m, G)\right] = o(m^{-1/2}), \tag{32}$$

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),

$$E\left[W_2^2(G_m, G)\right] \leq \frac{2}{m}\sum_{i=1}^{m} V(\eta_{d(i)}), \tag{33}$$

where $\eta_{d(1)} < ... < \eta_{d(m)}$ denotes the order statistic of an i.i.d. sample $(\eta_{d1}, ..., \eta_{dn})$ from G. Then, by Lemma 2, we have

$$\begin{aligned}\sum_{i=1}^{m} V(\eta_{d(i)}) &\lesssim E\left[\sum_{i=1}^{m} \frac{1}{i \wedge (n+1-i)}\left(\frac{1}{C^2} \vee \frac{\eta_{d(i)}^2 \ln(1+|\eta_{d(i)}|)^4}{K^2} + \eta_{d(i)}^2\right)\right] \\ &\lesssim \left(E[Z_m^2] + E[Z_m^2 \ln(1+Z_m)^4]\right) \sum_{i=1}^{\left[\frac{m+1}{2}\right]} \frac{1}{i} \\ &\lesssim E[Z_m^{2+\varepsilon/3}]\left[1 + \ln(m)\right], \end{aligned} \tag{34}$$

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

$$m^{-\frac{2+\varepsilon/3}{4+\varepsilon}} E[Z_m^{2+\varepsilon/3}] \leq \left\{m^{-1}E[Z_m^{4+\varepsilon}]\right\}^{\frac{2+\varepsilon/3}{4+\varepsilon}} = o(1), \tag{35}$$

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 (33), (34) and (35), we obtain (32).

B.5 Additional lemmas

The proof of Theorem 3 relies on two lemmas, which we state and prove below. Note that Lemma 2 is similar to Corollary 2.12 in Boucheron and Thomas (2015).

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 that $(T_1, ..., T_n)$ is an i.i.d. sample with marginal cdf F, survival function S and a positive density f. Then, for all $i \in \{1, ..., n\}$,*

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

B.5.1 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).

B.5.2 Proof of Lemma 2

First, note that

$$V(T_{(i)}) \leq 2\left[V(T_{(i)}F(T_{(i)})) + V(T_{(i)}S(T_{(i)}))\right]. \tag{36}$$

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,

$$\begin{aligned} V(T_{(i)}F(T_{(i)})) \leq & 4E\left[\sum_{k=n+1-i}^{n}\frac{\partial g}{\partial x_k}(E_{n+1-i},...,E_n)^2\right] \\ =& 4\left[\sum_{j=n+1-i}^{n}\frac{1}{j^2}\right]E\left[\left(\frac{F(T_{(i)})}{h(T_{(i)})}+S(T_{(i)})T_{(i)}\right)^2\right] \\ \leq & \frac{16}{n+1-i}E\left[\left(\frac{F(T_{(i)})S(T_{(i)})}{f(T_{(i)})}\right)^2+S(T_{(i)})^2T_{(i)}^2\right]. \end{aligned} \tag{37}$$

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:

$$V(T_{(i)}S(T_{(i)})) \leq \frac{16}{i}E\left[\left(\frac{F(T_{(i)})S(T_{(i)})}{f(T_{(i)})}\right)^2+F(T_{(i)})^2T_{(i)}^2\right]. \tag{38}$$

By combining (36), (37), (38) and $x^2+(1-x)^2 \leq 1$ for $0 \leq x \leq 1$, we finally obtain

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