

Two-Sample Least Squares Projection

David Pacini*

University of Bristol

February 15, 2016

Abstract

This paper investigates the problem of making inference on the coefficients in the linear projection $y = x'\alpha + z'\beta + u$ of an outcome variable y on covariates (z, x) when data are available from two independent random samples; the first sample contains information on only the variables (y, z) , while the second sample contains information on only the covariates. In this context, the validity of existing inference procedures depends crucially on the assumptions imposed on the joint distribution of (y, z, x) . This paper introduces a novel characterization of the identified set of the coefficients of interest when no assumption (except for the existence of second moments) on this joint distribution is imposed. One finding is that inference is necessary nonstandard because the function characterizing the identified set is a nondifferentiable (yet Bouligand directionally differentiable) function of the data. The paper then introduces a bias-corrected estimator and a nonconservative locally uniformly asymptotically valid confidence interval for the coefficients of interest based on the Bouligand directional differential. Monte Carlo experiments explore the numerical performance of the proposed estimator and confidence interval. Potential applications are discussed.

KEYWORDS: Identification; Least Squares Projection; Data Combination.

JEL CLASSIFICATION: C21, C26

*Author address: Department of Economics, University of Bristol, 8 Woodland Road, Bristol BS8 1TN, UK; Email: David.Pacini@bristol.ac.uk; Tel.: +44 (0) 11 79 28 84 37

1. INTRODUCTION

Least squares (or linear) projections coefficients are employed to approximate conditional expectations while guarding against misspecification and the curse of dimensionality (see Goldberger, 1991; Hayashi, 2001; or Wooldridge, 2002, for a textbook exposition). Economists who use survey data for making inference about these coefficients often face the situation when no single sample includes all the variables of interest but there are two independent samples and each variable is included in at least one.¹ In this context, complications arise because the coefficients of interest depend on moments of variables that are not jointly observed. The prominent method adopted to sidestep these complications is to impose additional assumptions on the distribution of the variables of interest. Assumptions include either restricting the dependence between the variables observed in different samples (see e.g., Rodgers, 1984) or requiring the presence of an instrumental variable observed in all samples (see e.g., Klevermarken, 1982). These assumptions are not testable. If there is doubt about their validity, then is worthy of consideration to analyze the sensitivity of inferences to a failure of them. Little is known, however, about how to make inference when no assumption on the joint distribution of the variables of interest is imposed, except that the coefficients of interest are not point identified (see e.g., Ridder and Moffitt, 2007).

Motivated by the previous situation, we study the problem of making inference on the coefficients α and β in the linear projection $y = x'\alpha + z'\beta + u$ of an outcome variable y on covariates (z, x) when data are available from two independent random samples; the first sample gives information on only (y, z) , while the second sample gives information on only the covariates. The disturbance term u is assumed to be uncorrelated with the covariates, and no assumption on the joint distribution of (y, x, z) , except for the existence of second moments, is imposed. We show that the identified set, that is, the set of values of the coefficients of interest compatible with knowledge of the distributions of (y, z) and of (z, x) , can be written as the intersection of two sets. We then derive a function characterizing the boundary of the identified

¹Examples include Japelli, Pischke and Souleles (1998); Meghir and Palme (1999); Carroll, Dynan and Krane (2003); Fang, Keane and Silverman (2008); Bostic, Stuart and Painter (2009); and Brzozowski, Gervais, Klein, and Suzuki (2010). In particular, consumption and wealth are seldom measured together for a single sample of households. Separate measurements however are more often available. In the US, for instance, the Consumer Expenditure Survey measures consumption and other household socioeconomic characteristics but wealth for a sample of households. The Survey of Consumer Finances measures wealth and socioeconomic characteristics but consumption for a different sample of households. Additional examples of this situation are discussed in the text (see Section 2.1) and in the survey papers by Chesher and Nesheim (2006) and Ridder and Moffitt (2007).

set, and use this function to construct an estimator and a confidence interval for the coefficients of interest. To our knowledge, the estimator and the confidence interval are new.

The construction of estimation and inference procedures is complicated because the function characterizing of the identified set involves the composition of max and min operations. It is well-known (see e.g., Hirano and Porter, 2012) that max and min operations render analog estimators systematically biased and invalidate the use of standard tools for inference (e.g., normal or nonparametric bootstrap approximation of sampling distributions). To overcome these difficulties, we construct an estimator which corrects the precision of the analog estimator of the identified set. The precision-correction is similar to the general nonparametric bootstrap bias correction but it is based on a version of the bootstrap that is different from the nonparametric one. This version of the bootstrap, which is a specific case of the modified version of the bootstrap by Fang and Santos (2014), is introduced to overcome the inconsistency of the nonparametric bootstrap in our context. The confidence interval is constructed by inverting a test statistic. Both the estimator and the confidence interval are based on an approximation to the directional differential of the function characterizing the identified set. The theoretical properties of the estimator and the confidence interval are discussed. Monte Carlo experiments illustrate the implementation and evaluate the numerical properties of the proposed procedures.

1.1. Related Literature

The problem of making inference on least squares projections from two independent samples has been studied in several strands of literature under different concerns and methodologies. We next review these strands of literature.

A first strand of related literature focuses on matching-based estimation of linear regression coefficients (see e.g., Ressler, 2002; D’Orazio, DiZio and Scanu, 2006). In this strand of literature, complications arising from the lack of observations on (y, x) are sidestepped by either imputing the values of covariates in the first sample (or the values of the outcome variable in the second sample) or by assuming that the conditional expectation of the outcome variable given the covariates is linear (see e.g., Rodgers, 1984; Rubin, 1986; Moriarity and Scheuren, 2003). The imputation procedures are valid under the assumption that the outcome variable and the covariates observed only in one sample are independent conditional on the common covariates.

This conditional independence assumption, or the linear conditional expectation assumption, however, often finds little justification in Econometric practice. We do not assume either that y and x are independent conditional on z or that the conditional expectation of the outcome variable given the covariates is linear. Our approach is thus useful to see what is lost when the latter conditional independence and/or the linearity assumption are not valid.

A second strand of related literature studies estimation and inference when instrumental variables are available (see e.g., Angrist and Krueger, 1992; Arellano and Meghir, 1992; Inoue and Solon, 2010). In this strand of literature, complications arising from the lack of observations on (y, x) are overcome by assuming that some of the variables common to both samples are instrumental variables. Although this instrumental variable assumption does deliver point identification, it is often the case that instrumental variables are not available. In this context, our results are useful to estimate least squares projection coefficients when there are no instrumental variables observed in both samples.

The last strand of related literature focuses on nonparametric identification of the conditional expectation of the outcome variable given the covariates when the common covariates are discrete (e.g., Vitale, 1979; Cross and Manski, 2002; Molinari and Peski, 2006). In this strand of literature, identification analysis is carried out without imposing additional assumptions delivering point identification. Our work is in the same spirit, but applies to a different setting. First, we do not restrict the common variables to be discrete. Second, our focus is on least squares linear projections rather than the conditional expectation itself. Third, our focus is on identification and inference and not just identification. Bontemps, Magnac and Maurin (2009) study identification of least squares linear predictors from two independent samples. We consider the same setup but our characterization of the identified set is different from theirs. As we discuss below, our characterization does make use of the marginal distributions of (y, z, x) while their characterization does not.

To close the review of the literature, we mention the differences between our problem and other problems studied in the literature on sample combination. The assumption that the two samples are independent distinguishes our problem from the problem with samples with common observational units (e.g., Devereux and Tripathi, 2009; Komarova, Nekipelov, Yakovlev, 2012; Poirer and Ziebarth, 2014). The fact that the two independent samples do not deliver point

identification distinguishes our paper from the literature studying either the case when the two samples jointly deliver point identification (see e.g., Chen, Hong and Tamer, 2005; Hirakawa and Prokhorov, 2014) or the case when one sample alone delivers point identification and a second sample is used for efficiency gains (see e.g., Hellerstein and Imbens, 1999). Fan, Sherman and Shum (2014) consider the related problem of combining samples to identify distributional treatment effects.

1.2. Organization of the Paper and Notation

The outline of the paper is as follows. In the next section, we define the coefficients of interest, describe the data, and discuss applications fitting our setup. In Section 3, we present the main result of the paper, namely an operational characterization of the identified set. We also discuss the force of additional assumptions to shrink the identified set to a singleton. In Section 4, we describe an estimator for the coefficients of interest and discuss the construction of a confidence interval. In Section 5, we explore via Monte Carlo exercises the finite-sample performance of the proposed estimator. Section 6 concludes. We collect the proofs of the results in appendices.

Through out, we use the subscript or superscript "o" to distinguish the "true" value of an unknown quantity from any other possible value that quantity may take. We consider a collection of observational units (i.e., individuals, households, etc.) to be studied at a given period in time and index an observational unit in this collection by i . For each i , we define the random vector (y_i, x'_i, z'_i) on a probability space with probability measure P_o . We suppress the subscript i in the notation whenever this can be done without causing confusion. The outcome variable y is univariate and the covariates (x', z') are random vectors of dimension d_x and d_z , respectively. We use the expression E to denote the expectation associated to P_o . We let $s_{xy}^o := E(xy)$ denote the value of the expectation of the product of y and x . Similarly, we define $s_{xx'}^o := E(xx')$, $s_{zz'}^o := E(zz')$, etc.. We denote the d -dimensional Euclidean space by \mathbb{R}^d and equip it with the Euclidean norm $\|\cdot\|$.

2. THE SETUP

We begin by describing the parameter of interest.

Assumption P (Parameter of Interest). Knowledge is sought about the coefficients $\theta_o = (\alpha'_o, \beta'_o)'$ defined by:

$$(P.i) \quad y = x'\alpha_o + z'\beta_o + u \quad \text{with} \quad E((x', z')'u) = 0,$$

where the joint distribution of (y, x', z') induced by P_o , say $F_{y|xz}^o$, is such that:

(P.ii) The variance of (y, x', z') is finite and positive semidefinite.

An equivalent way of writing Assumption (P.i) is $\theta_o := \arg \min_{(\alpha, \beta)} E[E(y|z, x) - x'\alpha - z'\beta]^2$, which shows that θ_o can be interpreted as the coefficients in the least squares projection of the conditional expectation of y given (x', z') under quadratic loss. Assumption (P.i) is weaker than the mean-independence restriction $E(y - x'\alpha_o - z'\beta_o|x, z) = 0$. Indeed, Assumption (P.i) does not restrict the conditional expectation of y given (x, z) to be a linear function. In empirical work, the difference between Assumption (P.i) and the mean-independence restriction is often overlooked. In our context however, this distinction is of importance because lack of correlation and mean-independence deliver different identification results. Assumption (P.ii) ensures enough variation to define θ_o .

If a sample with replications of the triplet (y, x', z') was available, inference on θ_o would be straightforward. Here we focus on the case when replications of this triplet are unavailable. We assume instead that data are available from two samples without common observational units:

Assumption D (Data). Let $y, z \mapsto G_{yz}^o(y, z)$ denote the (y, z) -marginal distribution of $y, x, z \mapsto F_{y|xz}^o(y, x, z)$. A similar notation is adopted for $x, z \mapsto G_{xz}^o(x, z)$. Data are available from two independent samples. The first sample $\{y_i, z'_i\}_{i=1}^{n_A}$ contains independent and identically distributed (iid) replications of the variables (y, z') generated from G_{yz}^o for a group of n_A observational units. The second sample $\{x'_j, z'_j\}_{j=n_A+1}^n$ contains iid replications of the variables (x', z') generated from G_{xz}^o for a group of different $n_B = n - n_A$ observational units.

2.1. Potential Applications

To illustrate the applicability of our setup, we now discuss potential applications fitting it. The first potential application comes from the work by Bostic, Gabriel, and Painter (2009, BGP from now on). They employ two-sample data to measure the dependence between log household consumption and log housing wealth after controlling for households characteristics. The standard specification proposed by BGP is $y_i = x'_i\alpha_o + z'_i\beta_o + u_i$, where y_i denote the log

consumption of household i living in the US in 2001, x_i is the log of housing wealth, z_i is a vector of household characteristics including income and household size and u_i is a disturbance term uncorrelated with z_i and x_i as in Assumption P. Interest is on α_o and one would ideally require information on (y, x', z') for a single sample of households. Since such data are not available, BGP employ data from two samples; the Consumer Expenditure Survey (CEX) and the Survey of Consumer Finances (SCF).² The CEX provides information on households' consumption and demographic characteristics only. The SCF in turn provides information only on demographic characteristics and housing wealth. The CEX and SCF do not survey the same households because they are independent samples. To overcome the lack of joint observations on (y, x) , BGP employ an imputation procedure. This procedure is valid under the assumption that log consumption and log housing wealth are independent conditional on household characteristics. We may be concerned with the situation when this conditional independence assumption does not hold. In this context, the results derived below are intended to evaluate the sensitivity of inferences about α_o to a failure of the conditional independence assumption.

The second potential application concerns the measurement of returns to education. Let y_i denote log hourly wages for a worker i , z_i a vector of worker characteristics including education and experience, and let x_i denote a proxy for worker ability such as the intelligence quotient (IQ) test score. Following the literature, one may decide to measure returns to education using the component in β_o associated to education. Only a few datasets contain measures of wages, education, experience and IQ test scores for a single sample of workers. However, these variables are measured in different samples from the same population. On one hand, household surveys carried out by government agencies usually gather information on wages, education and experience but not IQ test scores. On the other hand, there are surveys carried out by psychometricians gathering information on education and IQ test scores but wages and experience. In this context, the results derived below could be used to combine data from independent surveys to mitigate the undesirable consequences of omitted variables on estimates of the component of interest in β_o .

The third potential application comes from the marketing literature. To design marketing

²One important caveat to this example should be kept in mind. A key presumption underlying our analysis is that data are obtained by simple random sampling (see Assumption D). The SCF does not use a simple random but a dual frame sampling design. The identification results below however still apply.

campaigns, firms would like to infer the association, as measured by α_o , between the units y_i of a good purchased by consumer i and consumer's time exposure to advertising x_i . Collecting information on purchases and ads exposure for a single survey of consumers would be valuable but it is usually a very expensive proposition (see The Nielsen Company, 2009; Ipsos MORI, 2011). The common alternative among firms is to have access to independent samples as characterized in Assumption D. A first sample gathers information on purchases and consumers demographic characteristics z_i . A second independent sample contains information on ads exposure and the same consumers demographic characteristics.

3. IDENTIFICATION

In this section, we first define the *identified set* and describe the identification problem. We then introduce an operational characterization of this set. This characterization is the main result of the paper. We also discuss the force of additional assumptions to shrink the identified set to a singleton.

3.1. The Identification Problem

For identification purposes, we assume that the distributions $y, z \mapsto G_{yz}^o(y, z)$ and $x, z \mapsto G_{xz}^o(x, z)$ are known. For latter use, let define $d_o := -[s_{zz'}^o - s_{zx'}^o(s_{xx'}^o)^{-1}s_{xz'}^o]^{-1}s_{zx'}^o(s_{xx'}^o)^{-1}$, $c_o := [s_{zz'}^o - s_{zx'}^o(s_{xx'}^o)^{-1}s_{xz'}^o]^{-1}s_{zy}^o$, $b_o := -(s_{xx'}^o)^{-1}s_{xz'}^o c_o$ and $a_o := (s_{xx'}^o)^{-1} - (s_{xx'}^o)^{-1}s_{xz'}^o d_o$. For a component x_k of x , let $x_k, y \mapsto F_{ky}^o(x_k, y)$ denote the joint distribution function of x_k and y . Using Assumption (P.i), we write θ_o as the composition “ \circ ” of two linear functions:

$$\theta_o := f \circ s(F_{1y}^o, \dots, F_{ky}^o, \dots, F_{dxy}^o) \quad (1)$$

where the first linear function is

$$F_{1y}, \dots, F_{dxy} \mapsto s(F_{1y}, \dots, F_{dxy}) := \left(\int x_1 y dF_{1y}, \dots, \int x_{d_x} y dF_{dxy} \right)'$$

and the second linear function is

$$\lambda \mapsto f(\lambda) := \begin{pmatrix} a_o + b_o \lambda \\ c_o + d_o \lambda \end{pmatrix}$$

All the expectations in Equation (1) are known except for $s_{xy}^o = s(F_{1y}^o, \dots, F_{d_x y}^o)$. Solving the identification problem involves to exploit the restrictions imposed by Assumptions P and D on s_{xy}^o to ultimately recover θ_o .

To study how Assumption P restricts θ_o , let Λ denote the set of values of the expectation of the product between y and x such that the variance of (y, x', z') is positive semidefinite. Define the set Θ_M of values of θ_o compatible with this variance restriction by:

$$\Theta_M := \{\theta \in \mathbb{R}^{d_z + d_x} : \theta = f(\lambda) \text{ for any } \lambda \in \Lambda\}$$

Assumption D in turn restricts the joint distribution of the outcome variable and the covariates to have marginal distributions equal to the distributions characterizing the two samples. Let \mathcal{F} denote the set of joint distributions for the pairs $(y, x_1), \dots, (y, x_{d_x})$ compatible with knowledge of $y, z \mapsto G_{yz}^o(y, z)$ and $x, z \mapsto G_{xz}^o(x, z)$. Define the set Θ_F of values of θ_o compatible with the marginal distributions by:

$$\Theta_F := \{\theta \in \mathbb{R}^{d_z + d_x} : \theta = f \circ s(F_{x_1 y}, \dots, F_{x_k y}, \dots, F_{x_{d_x} y}) \text{ for any } (F_{x_1 y}, \dots, F_{x_k y}, \dots, F_{x_{d_x} y}) \in \mathcal{F}\}$$

Intersecting Θ_M and Θ_F one has:

Definition 1 (*Identified Set*). *The identified set Θ_o of θ_o is $\Theta_o := \Theta_M \cap \Theta_F$.*

The identification problem is to derive an *operational* characterization of Θ_o . By *operational* we mean a characterization suggesting an estimation procedure for θ_o . To our knowledge, a solution to this problem is not readily available in the literature. A partial solution has been proposed by Bontemps, Magnac and Maurin (2009), who characterize a set different from Θ_o . The set in Bontemps, Magnac and Maurin (2009) is defined only by the restriction that the variance of (y, x', z') must be positive semidefinite. By contrast, Θ_o is defined not only by the restriction on the variance of (y, x', z') but also by the restriction that the (y, z') - and (x', z') -marginals of the joint distribution of (y, x', z') must be equal to the distributions characterizing the available data.

3.2. Solving the Identification Problem

To proceed, we obtain a first characterization of Θ_o in terms of *support functions*.³ This characterization is attractive because it boils the identification problem down to solving two mathematical programming problems. The set Λ is convex because it is defined by a quadratic inequality (see Lemma 4 below). The set \mathcal{F} is convex as well because it is defined by linear restrictions (see Lemma 5 below). The sets Θ_M and Θ_F are linear transformations of the convex sets Λ and \mathcal{F} , respectively, because $\lambda \mapsto f(\lambda)$ and $F_{1y}, \dots, F_{d_x y} \mapsto f \circ s(F_{1y}, \dots, F_{d_x y})$ are linear. Since convexity is preserved under linear transformations (see Hiriart-Urruty and Lemarechal, 2004, Proposition 1.2.4), Θ_F and Θ_M are convex. Furthermore, since convex sets are characterized by their support functions (see Hiriart-Urruty and Lemarechal, 2004, Theorem 2.2.2), and the intersection of convex sets is equal to the minimum of their support functions (see Rockafellar, 1970, Corollary 16.5.1), Θ_o can be rewritten as:

Lemma 1 (*Characterization of the Identified Set*). *Let q denote a vector of directions belonging to the unit sphere $\mathbb{S}^{d_z+d_x}$ in $\mathbb{R}^{d_z+d_x}$. Then,*

(i) *The set Θ_M is characterized by:*

$$\Theta_M = \{\theta \in \mathbb{R}^{d_z+d_x} : q'\theta \leq s_M(q) := \sup_{\lambda \in \Lambda} q'f(\lambda) \text{ for any } q \in \mathbb{S}^{d_z+d_x}\}$$

where $q \mapsto s_M(q)$ is the support function of Θ_M .

(ii) *The set Θ_F is characterized by:*

$$\Theta_F = \{\theta \in \mathbb{R}^{d_z+d_x} : q'\theta \leq s_F(q) := \sup_{(F_{1y}, \dots, F_{d_x y}) \in \mathcal{F}} q'f \circ s(F_{1y}, \dots, F_{d_x y}) \text{ for any } q \in \mathbb{S}^{d_z+d_x}\}$$

where $q \mapsto s_F(q)$ is the support function of Θ_F .

(iii) *The identified set Θ_o is characterized by:*

$$\Theta_o = \{\theta \in \mathbb{R}^{d_z+d_x} : q'\theta \leq m(q) := \inf_{t \in \{M, F\}} s_t(q) \text{ for any } q \in \mathbb{S}^{d_z+d_x}\} \quad (2)$$

where $q \mapsto m(q)$ is the support function of Θ_o .

³The support function of a convex set is equal to the signed distance of supporting hyperplanes of the set from the origin.

Lemma 1 describes Θ_o as the collection of vectors θ whose linear combination with the vector of directions q is smaller or equal than the minimum of the value functions associated to the programming problems $\sup_{\lambda \in \Lambda} q'f(\lambda)$ and $\sup_{(F_{1y}, \dots, F_{d_x y}) \in \mathcal{F}} q'f \circ s(F_{1y}, \dots, F_{d_x y})$. We next study the solution to these programming problems, first for the case when x is univariate and then for the multivariate case.

3.3. The Covariate x is Univariate

Assuming that x is univariate simplifies the exposition. To characterize Θ_M , start from the observation that the restriction on the variance of (y, x', z') holds if and only if the determinant of this variance is nonnegative. In the next lemma, we use this observation to show that Λ is an interval. We further characterize the endpoints of this interval, and use this characterization to solve the programming problem characterizing Θ_M .

Lemma 2 (Operational Characterization of Θ_M when $d_x = 1$). *Let Assumptions (P) and (D) hold with $d_x = 1$. Let q denote a vector belonging to the unit sphere in \mathbb{R}^{1+d_z} . Split q into $q = (q_\alpha, q'_\beta)'$, where q_α is a vector of dimension $d_x = 1$ and q_β is a vector of dimension d_z . Define $v_q^o := q'_\alpha a_o + q'_\beta c_o$ and $e_q^o := q'_\alpha b_o + q'_\beta d_o$. Let $V(y)$ and $V(x)$ denote the variance of y and x , respectively. Let define ρ_{zy}^o as the element-by-element correlation between z and y .⁴ Define similarly ρ_{zx}^o . Define the moments:*

$$\begin{aligned} \lambda_{Ml}^o &:= E(y)E(x) + [V(y)V(x)]^{1/2} \left[\rho_{zy}^o \rho_{zx}^o - \sqrt{(1 - \rho_{zx}^o \rho_{zx}^o)(1 - \rho_{zy}^o \rho_{zy}^o)} \right] \\ \lambda_{Mu}^o &:= E(y)E(x) + [V(y)V(x)]^{1/2} \left[\rho_{zy}^o \rho_{zx}^o + \sqrt{(1 - \rho_{zx}^o \rho_{zx}^o)(1 - \rho_{zy}^o \rho_{zy}^o)} \right] \end{aligned}$$

Then, the support function characterizing Θ_M is:

$$s_M(q) = \max_{r \in \{l, u\}} v_q^o + e_q^o \lambda_{Mr}^o \quad (3)$$

To derive an operational characterization of Θ_F , recall that $s_F(q) = \sup_{F_{xy} \in \mathcal{F}} q'f \circ s(F_{xy})$. This programming problem is a variant of the Kantorovich optimal transportation problem $\sup_{F_{xy} \in \mathcal{F}} s(F_{xy})$. The difference is in the transformation $q'f(\cdot)$ applied to the total cost function

⁴For a random vector z and a random variable y , the k -th component of the vector of correlation coefficients ρ_{zy} is $[E(z_k y) - E(z_k)E(y)] / (V(z_k)V(y))^{1/2}$.

$s(F_{xy})$. By exploiting existing closed-form solutions for the Kantorovich optimal transportation problem (see e.g., Rachev and Ruschendorf, 1998, Chapter 3), we obtain the following result:

Lemma 3 (*Operational Characterization of Θ_F when $d_x = 1$*). *Let Assumptions (P) and (D) hold with $d_x = 1$. Define v_q and e_q as in Lemma 2. Let $y, z \mapsto G_{y|z}^o(y|z)$ denote the conditional distribution of y given z and let $\tau, z \mapsto Q_{x|z}^o(\tau|z)$ denote the conditional quantile function of x given z . Define the moments:*

$$\lambda_{Fl}^o := \mathbb{E}\left[yQ_{x|z}^o(1 - G_{y|z}^o(y|z)|z)\right] \quad , \quad \lambda_{Fu}^o := \mathbb{E}\left[yQ_{x|z}^o(G_{y|z}^o(y|z)|z)\right]$$

Then, the support function characterizing Θ_F is:

$$s_F(q) = \max_{r \in \{l, u\}} v_q^o + e_q^o \lambda_{Fr}^o \quad (4)$$

By intersecting the support functions in Lemma 2 and Lemma 3 according to the formula in Lemma 1, one has the following result:

Proposition 1 (*Operational Characterization of the Identified Set when $d_x = 1$*). *Let Assumptions (P) and (D) hold with $d_x = 1$. Let define v_q^o , e_q^o , λ_{tr}^o for $t \in \{M, F\}$ and $r \in \{l, u\}$, as in Lemmas 2 and 3. Then, the support function characterizing Θ_o is*

$$m(q) = \min_{t \in \{M, F\}} \max_{r \in \{l, u\}} v_q^o + e_q^o \lambda_{tr}^o \quad (5)$$

3.4. The Covariate x is Multivariate

We now derive a characterization of the identified set when x may be multivariate. Extending the characterization of s_M in Lemma 2 from the univariate to the multivariate case requires some elaboration because Λ is not longer an interval but an ellipsoid. The following Lemma uses the projection of y on z and of x on z to obtain this extension.

Lemma 4 (*Operational Characterization of Θ_M*). *Let Assumptions (P) and (D) hold. Define*

v_q^o and e_q^o as in Lemma 2 with the corresponding change in dimension to accommodate the case $d_x \geq 1$. Define the projection of y on z and the projection of x on z by $y = z'\delta + \sigma w_A$ and $x = \Pi z + \Sigma w_B$, respectively, where $\delta := E(zz')^{-1}E(zy)$, σ is the standard deviation of the residual $y - z'\delta$, w_A is a unit variance random variable, $\Pi := E(zz')^{-1}E(zx')$, Σ is the standard deviation of the residual $x - \Pi z$ and w_B is a unit variance random vector of dimension $d_x \times 1$. Define further $A_o := (\sigma\Sigma)^{-1}$ and $B_o := \Pi E(zy) - E(xz')\delta + \Pi E(zz')\delta$. Then, the support function characterizing Θ_M is:

$$s_M(q) = v_q^o + e_q^o B_o + (e_q^o A_o^{-2} e_q^o)^{1/2} \quad (6)$$

We now turn our attention to s_F . The strategy exploited to characterize this function when x is univariate carries over to the multivariate case. In particular, the linearity of the objective function $F_{1y}, \dots, F_{d_x y} \mapsto f \circ s(F_{1y}, \dots, F_{d_x y})$ allows to extend Lemma 3 by applying the method of proof there to each of the components of $f \circ s(F_{1y}, \dots, F_{d_x y})$:

Lemma 5 (Operational Characterization of Θ_F). *Let Assumptions (P) and (D) hold. Define v_q^o and e_q^o as in Lemma 2 with the corresponding change in dimension to accommodate the case $d_x \geq 1$. Let $e_{q,k}^o$ denote the k -th element of the vector e_q^o . Define $y, z \mapsto G_{y|z}^o(y|z)$ as in Lemma 3 and let $\tau, z \mapsto Q_{k|z}^o(\tau|z)$ denote the conditional quantile function of x_k given z . Define the moments:*

$$\lambda_{Fkl}^o := \mathbb{E}[y Q_{k|z}^o(1 - G_{y|z}^o(y|z)|z)], \quad \lambda_{Fku}^o := \mathbb{E}[y Q_{k|z}^o(G_{y|z}^o(y|z)|z)]$$

Then, the support function characterizing Θ_F is:

$$s_F(q) = \sum_{k=1}^{d_x} \max_{r \in \{l,u\}} v_q^o + e_{q,k}^o \lambda_{Fkr}^o \quad (7)$$

By combining Lemmas 4 and 5, one obtains the main result of the paper:

Theorem 1 (Operational Characterization of the Identified Set). *Let Assumptions (P) and (D)*

hold. Then, Θ_o is characterized by the support function:

$$m(q) = \min \left(v_q^o + e_q^o B_o + (e_q^o A_o^{-2} e_q^{o'})^{1/2}, \sum_{k=1}^{d_x} \max_{r \in \{l, u\}} v_q^o + e_{q,k}^o \lambda_{Fkr}^o \right) \quad (8)$$

where v_q^o , e_q^o , λ_{Fkl}^o , λ_{Fku}^o , A and B are defined as in Lemmas 4 and 5.

Remark 1. The restriction on the marginal distributions of the joint distribution of (y, x', z') does not imply the positive semidefinite restriction on the variance of (y, x', z') . This is because the latter is a restriction holding across the marginal distributions while the former is a restriction on a given marginal distribution. For some particular distributions of the outcome variable and the covariates (i.e., joint distribution of (y, x, z) is normal), we may have that Θ_M is a proper subset of Θ_F (i.e., $s_F(q) > s_M(q)$ for all q). This may give the impression that considering the intersection is unnecessary. This impression, however, is wrong because in general Θ_M is not a superset of Θ_F or Θ_M may be estimated more precisely than Θ_F .

3.5. Obtaining Point Identification

We have emphasized the fact that the maintained assumptions deliver set identification of θ_o . In a particular application, the identified set may be too wide to provide the desired information about θ_o . For this reason, we now discuss the force of additional assumptions to achieve point identification. For the sake of exposition, we focus on the case when x is univariate.

Let us dispose first of one restriction whose implications are immediate. If z and x are uncorrelated, β_o is point-identified and α_o is not. This result corresponds to the equality between the "short regression" and the "long regression" as discussed by Goldberger (1991) or the absence of omitted variable problem. This suggests that the two samples may be informative about (at least the sign of) β_o when the correlation between the covariates is small. We explore this point in the Monte Carlo experiments below.

If at least one of the elements in β_o is zero, then α_o is point identified. This is equivalent to assume that one of the common covariates is an instrumental variable (see e.g. Angrist and Krueger, 1992, Arellano and Meghir, 1992). To see why, fix z to be a scalar. Then, when $\beta_o = 0$, it follows from the identifying mapping that $\alpha_o = s_{yz}^o / s_{xz}^o$. In such a case, α_o is point identified because s_{yz}^o and s_{xz}^o so are. Inference in this case has been discussed by Inoue and Solon (2010) under a conditional homoscedasticity restriction on the disturbance term.

If the uncorrelatedness restriction in (P.i) is replaced by the mean-independence restriction $E[(y - x'\alpha_o - z'\beta_o)|x, z] = 0$, then θ_o is point identified. This is because the mean-independence restriction implies that any measurable function of z , such as z_k^2 , is uncorrelated with the disturbance term u . In such a case, any of these functions can be used as an instrument to point identify θ_o . After restricting z to be discrete, it also fits in the setting studied by Cross and Manski (2002) and Molinari and Peski (2006). Inference in this case has been discussed by Ichimura and Martinez-Sanchis (2010).

Finally, if y is conditionally independent of x given z , then θ_o is also point identified. Under this conditional independence assumption, $E(yx)$ is equal to $E[E(y|z)E(x|z)]$. Point identification follows after evaluating the identifying mapping at $E[E(y|z)E(x|z)]$. When z is a discrete variable, this assumption also justifies the procedure of interpreting the coefficient in the least squares linear projection of $E(y|z)$ onto $E(x|z)$ as α_o .

4. ESTIMATION AND INFERENCE

To reflect sampling variability, we now drop the assumption that the distributions of the two samples are known. We estimate these distributions and employ the characterization of the identified set in Theorem 1 to construct an estimator of the components of θ_o . We also discuss the construction of a confidence interval.

4.1. Estimand

We begin by describing the object to be estimated. Motivated by the applications discussed in Section 2, we are interested in the components of θ_o rather than in θ_o itself. We then estimate the one-dimensional projection of the identified set over the k -axis. Since the identified set is convex (see Lemma 1), this one-dimensional projection is an interval and its endpoints can be characterized using the support function in Theorem 1. For a given direction q , let define the vector of nuisance parameters

$$\eta_q^o := (v_q^o, e_q^o B_o + (e_q^o A_o^{-2} e_q^{o'})^{1/2}, e_{q,1}^o \lambda_{F1l}^o, e_{q,1}^o \lambda_{F1u}^o, \dots, e_{q,d_x}^o \lambda_{Fd_x l}^o, e_{q,d_x}^o \lambda_{Fd_x u}^o)'$$

Let q_l denote the k -negative canonical direction (i.e., q_l is a vector taking value -1 in position k and zero elsewhere) and let $q_u = -q_l$ denote the k -positive canonical direction. For any possible

value η_{q_b} of $\eta_{q_b}^o$ with $b \in \{l, u\}$, let use the support function in Theorem 1 to define the bounding functions

$$\begin{aligned} m_l(\eta_l) &:= -m(q_l) = -\min \left(v_{q_l} + e_{q_l}B + (e_{q_l}A^{-2}e'_{q_l})^{1/2}, \sum_{k=1}^{d_x} \max_{r \in \{l, u\}} v_{q_l} + e_{q_l, k} \lambda_{Fkr} \right) \\ m_u(\eta_u) &:= m(q_u) = \min \left(v_{q_u} + e_{q_u}B + (e_{q_u}A^{-2}e'_{q_u})^{1/2}, \sum_{k=1}^{d_x} \max_{r \in \{l, u\}} v_{q_u} + e_{q_u, k} \lambda_{Fkr} \right) \end{aligned}$$

where $\eta_b = \eta_{q_b}$ to save on notation. Since $q \mapsto m(q)$ describes the signed distance of supporting hyperplanes of the identified set from the origin, one has that the interval

$$[\theta_k] = [m_l(\eta_l^o), m_u(\eta_u^o)]$$

is the one-dimensional projection of the identified set over the k -axis.

4.2. Estimator

A natural idea to estimate $[\theta_k]$ would be to employ the sample analog principle. This approach however may systematically under(over)-estimate the true value of the endpoints of $[\theta_k]$. This is due to the presence of compositions of the min and max operators. With the aim of improving over the sample analog estimator, we introduce a bias-corrected estimator. This estimator has three steps. In the first step, we estimate the conditional distribution function $y, z \mapsto G_{y|z}^o(y|z)$ and the conditional quantile function $\tau, z \mapsto Q_{x|z}^o(\tau|z)$ by a nonparametric method. In the second step, we estimate v_q, e_q, λ_{Fr} for $r \in \{l, u\}$, B and A by their sample analogs. We denote these estimates $\hat{v}_q, \hat{e}_q, \hat{\lambda}_{Fr}$, etc. In the third step, a bias-correction term is subtracted to the sample analog estimator of the endpoints:

$$[\hat{\theta}_k] := \left[m_l(\hat{\eta}_l) - \kappa_l \hat{b}_l, m_u(\hat{\eta}_u) - \kappa_u \hat{b}_u \right]$$

where κ_l and κ_u are constants between zero and one, and \hat{b}_l and \hat{b}_u are estimates of the bias of the endpoints of the sample analog estimator (to be described below). The constants k_l and k_u are included to control the amount of bias-correction and to avoid highly variable estimates. When $\kappa_l = 0$ and $\kappa_u = 0$, no bias-adjustment is attempted and $[\hat{\theta}_k]$ is just the sample analog estimator of $[\theta_k]$. In this extreme case, no improvement can be expected. When $\kappa_l = 1$ and

$\kappa_u = 1$, a full bias-adjustment is attempted. In this other extreme case, bias improvements can be expected but the variance of the estimator may diverge to infinity (see the discussion below).

We now describe (\hat{b}_l, \hat{b}_u) . If $\eta_b \mapsto m_b(\eta_b)$ was differentiable, we could derive the sampling distribution of the sample analog estimator by using the Delta Method and the nonparametric bootstrap to consistently estimate its bias and correct for it. The function $\eta_b \mapsto m_b(\eta_b)$ however is nondifferentiable. This is due to the composition of *min* and *max* functions. This nondifferentiability, among other implications, rules out the possibility to use the nonparametric bootstrap to consistently estimate the bias of the sample analog estimator (see e.g., Hirano and Porter, 2012). However, $\eta_b \mapsto m_b(\eta_b)$ is Bouligand directional differentiable (see Appendix B). This allows to approximate the bias of the sample analog estimator using the following algorithm:

Algorithm 1 (Step-by-Step Calculation of Bias Correction)

Step 1 - Draw S pairs of bootstrap samples, say $\{(\{y_{is}, z_{is}\}_{i=1}^{n_A}, \{x_{js}, z_{js}\}_{j=n_A+1}^{n_B})\}$, for $s=1, \dots, S$, by resampling with replacement from the samples $\{y_i, z_i\}_{i=1}^{n_A}$ and $\{x_j, z_j\}_{j=n_A+1}^{n_B}$.

Step 2 - Let $\hat{\eta}_b$ denote the estimate of the vector of nuisance parameters η_b^0 . Let $\hat{\eta}_{bs}^*$ denote the estimate of the vector of nuisance parameters η_b^0 computed from the bootstrapped samples s . For each sample s , calculate $\hat{\xi}_{bs}^* := \hat{m}_b(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}_{bs}^* - \hat{\eta}_b))$ for $b \in l, u$, where \hat{m}_b is a consistent estimator of m_b .

Step 3 - The delta bootstrap estimate of the bias of the sample analog estimator is

$$\hat{b}_b := \frac{1}{S} \sum_{s=1}^S n_A^{-1/2} \hat{\xi}_{bs}^*$$

The next Theorem establishes sufficient conditions under which (\hat{b}_l, \hat{b}_u) is consistent.

Theorem 2 (Consistent Bias Estimation). *Let Assumptions P and D hold. Let \mathcal{Q} and \mathcal{G} denote the parameter spaces for $\tau, z \mapsto Q_k(\tau|z)$, for all $k = 1, \dots, d_x$, and $y, z \mapsto G(y|z)$, respectively. Equip these spaces with norms $\|\cdot\|_{\mathcal{Q}}$ and $\|\cdot\|_{\mathcal{G}}$. Let assume that there are nonparametric estimators \hat{Q}_k and \hat{G} of $\tau, z \mapsto Q_{k|z}^o(\tau|z)$ and $y, z \mapsto G_{y|z}^o(y|z)$, respectively. Let further assume:*

(C.1) $G_{y|z}^o \in \mathcal{G}$; $\hat{G} \in \mathcal{G}$ with probability tending to one; and, for any number $0 \leq \delta \leq 1/4$, $n_A^\delta \|\hat{G} - G_{y|z}^o\|_{\mathcal{G}} = o_P(1)$.

(C.2) For all $k = 1, \dots, d_x$, $Q_{k|z}^o \in \mathcal{Q}_k$; $\hat{Q}_k \in \mathcal{Q}_k$ with probability tending to one; and, for any number $0 \leq \delta \leq 1/4$, $n_B^\delta \|\hat{Q}_k - Q_{k|z}^o\|_{\mathcal{Q}} = o_P(1)$.

(C.3) For all z and k , the density $x \mapsto g_{k|z}^o(x_k|z)$ associated to $x \mapsto G_{k|z}^o(x|z)$ is bounded and bounded away from zero.

(C.4) For some functions $z_j \mapsto \varphi_Q(z_j)$ and $z_i \mapsto \varphi_G(z_i)$, $\hat{Q}_k(\tau|z) - Q_{k|z}^o(\tau|z)$ and $\hat{G}(y|z) - G_{y|z}^o(y|z)$ admit the asymptotic linear expansions:

$$\begin{aligned}\hat{Q}_k(\tau|z) - Q_{k|z}^o(\tau|z) &= n_B^{-1} \sum_{j=n_A+1}^n \frac{[\mathbf{1}(x_j \leq Q_{k|z}^o(\tau|z))\varphi_Q(z_j) - \tau]}{g_{x|z}^o(Q_{k|z}^o(\tau|z))} + R_{Q_n}(\tau, z) \\ \hat{G}(y|z) - G_{y|z}^o(y|z) &= n_A^{-1} \sum_{i=1}^{n_A} [\mathbf{1}(y_i \leq y)\varphi_G(z_i) - G_{y|z}^o(y|z)] + R_{G_n}(y, z)\end{aligned}$$

with $R_{Q_n}(\tau, z) = o_P(n_B^\delta)$ uniformly over (τ, z) and $R_{G_n}(y, z) = o_P(n_B^\delta)$ uniformly over (y, z) for some number $0 < \delta < 1/2$.

Let $(b_l, b_u)' := n_A^{1/2} E[m_l(\hat{\eta}_l) - m_l(\eta_l^o), m_u(\hat{\eta}_u) - m_u(\eta_u^o)]$ denote the asymptotic bias of the sample analog estimator $[m_l(\hat{\eta}_l), m_u(\hat{\eta}_u)]$. Define (\hat{b}_l, \hat{b}_u) as in Algorithm 1. Then,

$$\|n_A^{1/2}(\hat{b}_l, \hat{b}_u)' - (b_l, b_u)'\| = o_P(1)$$

Remark 2. Algorithm 1 approximates the asymptotic bias of the sample analog estimator by the Monte Carlo mean of the bootstrap quantity $\hat{m}_b(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}_{bs}^* - \hat{\eta}_b))$. The approximation (\hat{b}_l, \hat{b}_u) is different from either the nonparametric bootstrap, which is the Monte Carlo mean of the bootstrap quantity $n^{1/2}[m_b(\eta_{bs}^*) - m_b(\hat{\eta}_b)]$, or the naive plug-in approximation $\hat{m}_b(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}_{bs}^* - \hat{\eta}_b))$. Implementing Algorithm 1 requires a consistent estimator of the directional differential. Such an estimator is constructed in Appendix B.

The nondifferentiability of $\eta_b \mapsto m_b(\eta_b)$ has two further implications for the evaluation of the bias-adjusted estimator. First, impossibility results for nondifferentiable functions (see e.g., Hirano and Porter, 2012, Theorem 2) imply that potential reductions in bias may be offset by an increase in variance. This implication does not preclude modifying procedures to mitigate the imprecision problem, but suggests that one should assess carefully the properties of the modified procedure. In the next section, we evaluate this bias-variance trade-off for different choices of (κ_l, κ_u) via Monte Carlo exercises. Second, it implies that standard notions of asymptotic efficiency (i.e., variance bounds associated to minimum variance unbiased estimators) will not

lead to useful comparisons between different estimators. Given this situation, we rely again on Monte Carlo exercises to evaluate the performance of the proposed estimator.

4.3. Confidence Interval

Consider now the problem of inference. To communicate sampling variability, one may wish to construct a confidence interval for θ_{k_0} . Fulfilling this wish is a delicate issue because the nondifferentiability of the bounding functions are responsible for coverage failures associated to confidence intervals constructed using standard techniques such as uniform normal approximations, bootstrapping or subsampling the sample analogs of the endpoints of $[\theta_k]$. To deal with the issues raised by nondifferentiability, we consider the confidence interval:

$$C_n := \{\theta_k \in \mathbb{R} : T_n(\theta_k) \leq \hat{q}_{1-\tau}\}$$

where

$$T_n(\theta_k) := \max\left(n^{1/2}[m_l(\hat{\eta}_l) - \theta_k], 0\right)^2 + \min\left(n^{1/2}[m_u(\hat{\eta}_u) - \theta_k], 0\right)^2$$

and $\hat{q}_{1-\tau}$ is a simulated critical value computed according to the following algorithm:

Algorithm 2. (Step-by-Step Calculation of C_n)

Step 1. For $\hat{\xi}_s^*$ calculated as in Algorithm 1, calculate $\hat{T}_{ns}^* := \max(\hat{\xi}_{ls}^*, 0)^2 + \min(\hat{\xi}_{us}^*, 0)^2$.

Step 2. Fix $\tau \in (0, 1)$. Set $\hat{q}_{1-\tau}$ equal to the $1 - \tau$ empirical quantile of $\{\hat{T}_{ns}^*\}_{s=1}^S$.

Step 3. Create a grid in \mathbb{R} . Let θ_{kc} denote a point in this grid.

Step 4. Calculate $T_n(\theta_{kc})$ for each θ_{kc} in the grid. Accept θ_c in the confidence interval if $T_n(\theta_{kc}) < \hat{q}_{1-\tau}$ otherwise discard θ_{kc} . Take the smallest and largest accepted values as the endpoints of C_n .

The following Theorem establishes the validity of C_n :

Theorem 3 (Locally Uniform Asymptotic Confidence Interval). *Let Assumptions P, D and C.1-C.4 hold. Let \mathcal{H} denote the parameter space for $\eta_o = (\eta_l^o, \eta_u^o)$. Let K denote any compact*

set of \mathcal{H} containing η_o . Consider the $n^{1/2}$ -contingent cone at η_o with respect to K :

$$\mathcal{K} := \{d \in \mathcal{H} : \exists d_n \text{ satisfying } \lim_{n \rightarrow \infty} \|d_n - d\|_{\mathcal{H}} = 0, \eta_o + n^{-1/2}d_n \in K\}$$

and $n^{-1/2}$ -local perturbation neighborhood of $P_o \in \mathcal{P}$:

$$\mathcal{P}_o := \{P_{\eta_o + n^{-1/2}d} \in \mathcal{P} : d \in \mathcal{K}\}$$

Then,

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_o, \theta_k \in [\theta_k]} P(\theta_k \in C_n) = 1 - \tau$$

Theorem 3 guarantees that C_n asymptotically covers θ_{k_o} with at least the pre-specified level $1 - \tau$ for whatever value of the nuisance parameters associated to a probability function in \mathcal{P}_o and whatever value of θ_{k_o} in the one-dimensional projection $[\theta_k]$. This includes the case when θ_{k_o} is either point-identified (i.e., $m_l(\eta_l^o) = m_u(\eta_u^o) = \theta_{k_o}$ because x and z are uncorrelated or because $\theta_{k_o} = \alpha_o$ and some β_o is zero) or interval-identified (i.e., $m_l(\eta_l^o) < m_u(\eta_u^o)$).

Remark 3. Unlike existing confidence intervals for parameters defined by moments inequalities, (i) the calculation of the critical value involves the use of a consistent approximation of the directional differential of the bounding functions; (ii) C_n is locally uniformly valid over nuisance parameters entering as an argument of nonsmooth functions. Unlike existing confidence intervals for parameters defined by nonsmooth functions, C_n is not conservative. The compactness restriction on the perturbation d is imposed to circumvent the impossibility result in Hirano and Porter (2012, Theorem 2b).

5. MONTE CARLO EXPERIMENTS

In this section, we employ simulated data to evaluate the performance of the procedures described in the previous section. The experiments show that:

- (i) When the correlation between z and x is low, the two samples may be informative about the sign of α_o .
- (ii) When inference does not take the restrictions on the marginal distributions of the joint

distribution of the variables of interest into consideration, the estimated intervals do not take the best advantage of the data, resulting in estimates wider than necessary.

(iii) Naive confidence intervals (i.e., based on nonparametric bootstrapping sample analog of endpoints) suffer from coverage problems. For medium-large sample sizes (e.g., 500 - 1000 observations in each sample) the confidence intervals described in the previous section resolve these problems.

5.1. Design of Experiments

We consider two Monte Carlo designs. For computational simplicity, we let x_i and z_i to be univariate and include a constant. For the true coefficients $\theta_o := (\alpha_{1o}, \beta_{1o}, \beta_{2o}) = (0, 1, 1)$, we generate y_i according to $y_i = \beta_{1o} + z_i\beta_{2o} + x_i\alpha_{1o} + u_i$, where u_i is a standard normal random variable independent of the covariates. In the first design, the joint distribution of (z_i, x_i) is bivariate normal with mean zero and unit variance. We call this design the "Normal Experiment". The design variable in this experiment is the correlation ρ_{xz} between z_i and x_i . Recall from Section 3.1 that β_2 is point identified when x_i and z_i are uncorrelated, i.e., when $\rho_{xz} = 0$. We use different values of ρ_{xz} to evaluate the uniform properties of confidence intervals. In the second design, we simulate z_i from a log-normal distribution with log-scale 0 and shape parameter 1 and x_i from $x_i = \delta z_i + v_i$, where δ is a constant and v_i is a random variable with standard normal distribution. We call this design the "Non-Normal Experiment". The design variable in this experiment is δ , which controls the correlation between x_i and z_i . β_2 is point identified when $\delta = 0$. In order to create two independent samples, we split the n draws of (y, x, z) into two samples of size n_A and n_B , respectively. In the first sample, we drop the realized values of x . In the second sample, we drop the realized values of y . We choose $n_A = n_B \in \{250, 500, 1000\}$. The number of replications in each experiment is 250.⁵

5.2. Performance Measures for Post-Simulation Analysis

To evaluate the finite-sample properties of different interval estimators, we now describe some performance measures. To the best of our knowledge, there is no widely accepted loss

⁵We find increasing the number of replications computationally costly, especially for the largest sample size in consideration (i.e., $n_A = n_B = 1000$). For the smallest sample size (i.e., $n_A = n_B = 250$), we also report results for 1000 Monte Carlo replications (see Appendix E). The qualitative conclusions obtained from 250 replications are not affected.

function to evaluate interval estimators. Given this state of affairs, we decide to use as loss function the mean squared error uniformly integrated over an interval (MSEI). To describe the MSEI loss function, let $\tilde{\theta}_{ks}$, for $s \in \{l, u\}$, denote any estimator of the endpoints of the interval of interest $[\theta_k]$. For the interval estimator $[\tilde{\theta}_k] := [\tilde{\theta}_{kl}, \tilde{\theta}_{ku}]$, the mean squared error uniformly integrated (MSEI) is defined as:

$$MSEI([\tilde{\theta}_k]) := E \left(\int_{\tilde{\theta}_{kl}}^{\tilde{\theta}_{ku}} (\theta_k - \theta_{ko})^2 \frac{d\theta_k}{\tilde{\theta}_{ku} - \tilde{\theta}_{lu}} \right)$$

Magnac and Maurin (2008) show that $MSEI([\tilde{\theta}_k])$ has the following decomposition:

$$MSEI([\tilde{\theta}_k]) = \underbrace{(\bar{\theta}_k - \theta_{ko})^2}_{\text{Dec}} + \underbrace{\frac{1}{3} \left(\frac{\bar{\theta}_{ku} - \bar{\theta}_{kl}}{2} \right)^2}_{\text{AL}} + \underbrace{\frac{1}{3} E \left((\tilde{\theta}_{kl} - \theta_{ko})^2 + (\tilde{\theta}_{ku} - \theta_{ko})^2 + (\tilde{\theta}_{kl} - \theta_{ko})(\tilde{\theta}_{ku} - \theta_{ko}) \right)}_{\text{ASE}}$$

where $\bar{\theta}_{ks} := E(\tilde{\theta}_{ks})$ denotes the expected value of the estimator $\tilde{\theta}_{ks}$, for $s \in \{l, u\}$, and $\bar{\theta}_k := (\bar{\theta}_{kl} + \bar{\theta}_{ku})/2$ is the expected average of the estimated lower and upper bound. The first term (denoted Dec) can be interpreted as the square of the familiar bias term. The second term (AL) can be interpreted as the specific ambiguity due to set identification instead of point identification. The third term (ASE) can be interpreted as the usual variance term. As pointed out by Magnac and Maurin (2008), the latter decomposition is an adaptation of the usual decomposition of the mean squared error to the case when identification is partial.⁶

5.3. Comparison of Estimators

We calculate five different estimators. The first estimator, called *Matrix*, is the sample analog of the one-dimensional projection of the set implied by the positive semidefinite restriction on the variance of (y, x', z') . The second estimator, called *Frechet*, is the sample analog of the one-dimensional projection of the set implied by the marginal restrictions on the joint distribution of (y, x', z') . The third estimator, called *Analog*, is the sample analog of the one-dimensional projection of the identified set. The fourth and fifth estimators are biased-corrected estimators. One of them, denoted *Bias-C(.5)*, attempts a partial bias-correction by setting the constants κ_l

⁶An alternative to the MSEI is a loss function weighting coverage and length of the interval estimators. We are not aware, however, of the use of this type of loss functions in the context of set identifying models. For the sake of completeness, we report the Monte Carlo coverage and average length as well.

and κ_u 1 equal to one half. The other one, denoted *Bias-C(1)*, attempts a full bias-correction by setting the constants κ_l and κ_u to one. We set the correlation between x and z to .2

Table 1 compares the finite performance of these six estimators when data are generated according to the design in the Normal experiment. For all sample sizes, the analog estimator based only on the marginal restrictions (Frechet) improves, in terms of MSEI, upon the analog estimator estimator based on the positive semidefinite variance restrictions (Matrix). This improvement can be attributed to the shorter length of the Frechet estimator (see the row labeled 'AL' in the Table). The Analog estimator offers modest improvements upon the Frechet estimator. These improvements come again from a reduction in average length. The bias-corrected estimators Bias-C(.5) and Bias-C(1) improve upon the Analog estimator. Their decentering and average standard error are similar to those of the Analog estimator, but the average length is smaller. In particular, Bias-C(1) performs better than Bias-C(.5). Table 2 compares the finite performance of the estimators when data are generated according to the Non-Normal experiment. Bias-C(.5) and Bias-C(1) have a decentering and average length terms smaller than Analog but a larger average standard deviation. The variance of Bias-C(.5) and Bias-C(1) exceed as well that of Matrix. This suggests that the gains in terms of bias obtained from bias-correction in some cases may be offset by an increase in the variance, as predicted by the theory.

In both data generating processes the estimators recover, on average, the sign α_o . We may attribute this result to the low correlation between the covariates in the experiments. Moreover, Analog may improve upon Matrix. This suggests that ignoring the marginal restrictions on the joint distribution of the variables of interest may give up information about the parameters of interest. The Frechet, Analog and Bias-C estimators are obtained after replacing $y, z \mapsto G_{y|z}^o(y|z)$ and $\tau, z \mapsto Q_{x|z}^o(\tau|z)$ by nonparametric estimates. We estimate these functions by series of cubic splines. We choose the number of knots, say K_n and L_n , according to the rule $K_n = L_n = \lfloor n_A^{1/3} \rfloor$. The sensitivity of results to different number of knots is explored in Appendix D. Implementing the bias-corrected estimator requires to choose a tuning parameter for estimating the directional derivative. This parameter has to diverge to infinity when the sample size increases and has to converge to zero when multiplied by the inverse of the square root of the sample size (see our discussion in Appendix B). In the simulations, we set this

parameter to $\log(n_A)$.⁷

5.4. Comparison of Confidence Intervals

We now consider the estimation of confidence intervals. We implement the confidence interval in Algorithm 2 for a nominal confidence level equal to .95. We call this confidence interval Delta. To evaluate the behavior of the Delta confidence interval across different values of the nuisance parameters, we choose values for ρ_{xz} in $\{.4, .2, 0\}$. In the experiment with $\rho_{xz} = 0$, the data generating process delivers point-identification of β_2 . In the other two experiments ($\rho_{xz} = .4$ and $\rho_{xz} = .2$), the data generating process delivers only set-identification.

Implementing the Delta confidence intervals is computationally intensive. It is worth then to explore the properties of less computationally intensive alternatives. We construct percentile nonparametric bootstrap confidence intervals (called *Non-Parametric*). The lower endpoint of this confidence interval is the .025-quantile of the non-parametric bootstrap distribution of the sample analog estimator of the lower endpoint of the one-dimensional projection of the identified set. The upper endpoint is the .975-quantile of the non-parametric bootstrap distribution of the sample analog estimator of the upper endpoint of the one-dimensional projection of the identified set. We expect these naive confidence intervals to perform worse than Delta. We also implement a version of the confidence interval described in Chernozhukov, Lee and Rosen (2009). We call this confidence interval *Before*. We set the number of bootstrap replications to 250.

Table 3 presents the actual coverage probability and the average length of the Nonparametric, Before and Delta confidence intervals in the Normal experiment. Table 4 presents the actual coverage probability and the average length in the Non-Normal experiment. Results in both designs are qualitatively similar.

6. SUMMARY AND CONCLUSIONS

Applied researchers interested in making inference about least squares projection coefficients are often confronted to the situation when the relevant variables are measured in two or more independent samples, neither of which contains information on all the variables of interest. When

⁷All the experiments were carried out in the program R using the libraries "splines" (to generate cubic spline basis) and "quantreg" (to estimate $\tau, z \mapsto Q_{x|z}^o(\tau|z)$).

no additional assumptions on the data generating process are invoked, the existing literature has shown that the coefficients of interest are not point-identified. This paper characterizes the identified set for the coefficients of interest and introduces a bias-corrected estimator and a confidence interval for them. The proposed estimator and confidence interval exploit the fact that the function characterizing the identified set is Bouligand directional differentiable. The results highlight the importance of exploiting the structure of the function characterizing the identified set when constructing inference procedures.

There are at least two topics which deserve further research. The first topic relates to the choice of the smoothing parameters for the estimators of the identified set. The second topic concerns theoretical comparison of alternative estimation procedures for intervals with nondifferentiable endpoints.

Acknowledgments. *I wish to thank Thierry Magnac, Stephane Bonhomme, Christian Bon-temps, Andrew Chesher, Fabiana Gomez, Gregory Jolivet, Pascal Lavergne, Nour Meddahi, Jorge Ponce, Christoph Rothe, Senay Sokullu, Frank Windmeijer and seminar participants at the Workshop on Set Identified Models in Toulouse '13, University of Bristol, the European Winter Meeting of the Econometric Society '11, Banco Central del Uruguay, the EC² '10 meeting, Universidad Carlos III/Madrid, the NESG '10 meeting, the ENTER Jamboree '10 meeting, and Toulouse School of Economics have offered useful comments and suggestions for which I am very grateful. All remaining errors are my responsibility.*

TABLE 1. *Monte Carlo Experiments: Comparison of Different Estimators.*
Normal Experiment.

Obs.	Covariate z ($\beta_2 = 1$)						Covariate x ($\alpha_1 = 1$)								
	Matrix	Frechet	Analog	Bias-C(.5)	Bias-C(1)	Matrix	Frechet	Analog	Bias-C(.5)	Bias-C(1)	Matrix	Frechet	Analog	Bias-C(.5)	Bias-C(1)
250	Mean	[.911,1.49]	[.934,1.47]	[.935,1.47]	[.965,1.44]	[.996,1.41]	[-1.41,1.41]	[-1.29,1.30]	[-1.29,1.30]	[-1.27,1.28]	[-1.25,1.26]	.987			
	Dec	.042	.041	.041	.041	.042	1.00	.988	.990	.988	.987				
	AL	.028	.024	.024	.019	.014	.669	.563	.562	.544	.526				
	ASE	.018	.018	.018	.018	.018	.003	.005	.004	.005	.005				
	RMSEI	.299	.291	.291	.281	.274	1.29	1.24	1.24	1.24	1.23				
	Cove.	72%	68%	68%	60%	53%	100%	100%	100%	100%	99%				
	Length	.587	.538	.538	.478	.418	2.83	2.60	2.59	2.55	2.51				
500	Mean	[.925,1.49]	[.938,1.47]	[.938,1.47]	[.961,1.45]	[.984,1.43]	[-1.41,1.41]	[-1.35,1.35]	[-1.34,1.34]	[-1.33,1.33]	[-1.31,1.31]				
	Dec	.043	.043	.043	.043	.043	1.00	1.00	.999	1.00	1.00				
	AL	.026	.024	.024	.021	.016	.669	.608	.606	.592	.578				
	ASE	.008	.008	.008	.008	.008	.001	.002	.002	.002	.002				
	RMSEI	.281	.276	.276	.269	.263	1.29	1.26	1.26	1.26	1.25				
	Cove.	74%	72%	72%	65%	54%	100%	100%	100%	100%	100%				
	Length	.566	.541	.539	.495	.451	2.83	2.70	2.69	2.66	2.63				
1000	Mean	[.921,1.49]	[.931,1.48]	[.931,1.48]	[.947,1.46]	[.963,1.35]	[-1.41,1.41]	[-1.36,1.37]	[-1.36,1.36]	[-1.35,1.35]	[-1.33,1.34]				
	Dec	.043	.043	.043	.043	.043	1.00	.991	.992	.992	.992				
	AL	.027	.025	.025	.022	.021	.671	.622	.622	.611	.600				
	ASE	.004	.004	.004	.004	.004	.001	.001	.001	.001	.001				
	RMSEI	.274	.271	.271	.265	.261	1.29	1.27	1.27	1.26	1.26				
	Cove.	82%	80%	80%	75%	70%	100%	100%	100%	100%	100%				
	Length	.576	.555	.554	.522	.491	2.83	2.73	2.73	2.70	2.69				

This table presents different measures describing the finite sample performance of different estimators of the coefficients of in the Normal Experiment. All details about this experiments are in Section 5.1. We set the correlation between covariates equal to $\rho_{zx} = .2$ and the number of knots in the estimation of the conditional quantile and distributions functions according to $K_n = L_n = \lfloor n^{.5} \rfloor$. The label "Obs." indicates the number of observations in each sample. "Matrix" is the sample analog estimator based on the positive definite restriction of the variance of (y, x, z) . "Frechet" is the sample analog estimator based on the restrictions on the marginal distributions of (y, x, z) . "Analog" is the sample analog estimator based on both the positive definite restriction and the restrictions on the marginal distributions of (y, x, z) . "Bias-C(.5)" is the after-precision-adjusted estimator described in Section 4.2. with bias-adjustment term equal to one half. "Bias-C(1)" is the after-precision-adjusted estimator described in Section 4.2. with bias-adjustment term equal to one. "Before" is the before-precision-adjusted estimator described in Section 4.2. 'Dec' stands for decentering of the mid-point of the interval, 'AL' is the adjusted length of the interval, 'ASE' is the variance of the estimators of the bounds, and 'RMSEI' is the square root of the mean squared error uniformly integrated defined in Section 5.2. The number of Monte Carlo replications is 250.

TABLE 2. Monte Carlo Experiments: Comparison of Different Estimators.
Non-Normal Experiment.

Obs.	Covariate z ($\beta_2 = 1$)					Covariate x ($\alpha_1 = 1$)				
	Matrix	Frechet	Analog	Bias-C(.5)	Bias-C(1)	Matrix	Frechet	Analog	Bias-C(.5)	Bias-C(1)
250	[.939,1.21]	[.924,1.21]	[.944,1.20]	[.961,1.18]	[.978,1.17]	[-1.40,1.45]	[-1.42,1.57]	[-1.27,1.40]	[-1.29,1.39]	[-1.31,1.39]
Mean	.006	.004	.005	.005	.005	.947	.851	.880	.900	.921
Dec	.006	.006	.005	.004	.003	.678	.751	.598	.605	.612
AL	.115	.129	.120	.135	.152	.008	.184	.053	.054	.059
ASE	.358	.375	.362	.381	.401	1.27	1.33	1.23	1.24	1.26
RMSEI	29%	31%	28%	24%	20%	99%	98%	97%	97%	98%
Cove.	.277	.289	.257	.224	.192	2.85	3.00	2.68	2.69	2.71
Length										
500	[.960,1.22]	[.950,1.22]	[.963,1.21]	[.975,1.20]	[.987,1.19]	[-1.40,1.43]	[-1.43,1.52]	[-1.31,1.39]	[-1.33,1.39]	[-1.34,1.39]
Mean	.008	.007	.007	.007	.008	.972	.913	.925	.945	.935
Dec	.005	.006	.005	.004	.003	.672	.726	.612	.627	.620
AL	.063	.071	.066	.074	.084	.001	.064	.017	.020	.024
ASE	.278	.291	.282	.295	.309	1.28	1.30	1.24	1.25	1.26
RMSEI	40%	38%	37%	32%	27%	100%	99%	99%	99%	99%
Cove.	.264	.273	.251	.227	.204	2.84	2.95	2.71	2.72	2.74
Length										
1000	[.959,1.22]	[.952,1.22]	[.963,1.21]	[.973,1.21]	[.983,1.20]	[-1.41,1.42]	[-1.38,1.47]	[-1.31,1.37]	[-1.33,1.38]	[-1.34,1.38]
Mean	.008	.007	.008	.008	.008	.984	.913	.941	.950	.959
Dec	.006	.006	.005	.004	.003	.672	.682	.603	.613	.623
AL	.035	.041	.038	.041	.044	.0008	.055	.016	.019	.023
ASE	.225	.235	.228	.233	.239	1.28	1.28	1.24	1.25	1.26
RMSEI	50%	46%	46%	42%	38%	100%	99%	99%	99%	99%
Cove.	.269	.271	.254	.236	.218	2.84	2.86	2.69	2.71	2.73
Length										

This table presents different measures describing the finite sample performance of different estimators of the coefficients of in the Non-Normal Experiment. All details about this experiments are in Section 5.1. We set the correlation between covariates equal to $\rho_{zx} = .2$ and the number of knots in the estimation of the conditional quantile and distributions functions according to $K_n = L_n = \lfloor n^{.3} \rfloor$. The label "Obs." indicates the number of observations in each sample. "Matrix" is the sample analog estimator based on the positive definite restriction of the variance of (y, x, z) . "Frechet" is the sample analog estimator based on the restrictions on the marginal distributions of (y, x, z) . "Analog" is the sample analog estimator based on both the positive definite restriction and the restrictions on the marginal distributions of (y, x, z) . "Bias-C(.5)" is the after-precision-adjusted estimator described in Section 4.2. with bias-adjustment term equal to one half. "Bias-C(1)" is the after-precision-adjusted estimator described in Section 4.2. with bias-adjustment term equal to one. "Before" is the before-precision-adjusted estimator described in Section 4.2. "Dec" stands for decentering of the mid-point of the interval, "AL" is the adjusted length of the interval, "ASE" is the variance of the estimators of the bounds, and "RMSEI" is the square root of the mean squared error uniformly integrated defined in Section 5.2. The number of Monte Carlo replications is 250.

TABLE 3. *Monte Carlo Experiments: Performance of Confidence Intervals.*

Normal Experiment.

Obs.	Procedure		Covariate z ($\beta_2 = 1$)			Covariate x ($\alpha_1 = 1$)		
			.4	.2	0	.4	.2	0
250	Nonparametric	Coverage	100%	100%	100%	100%	100%	100%
		Length	4.63	3.49	2.62	2.70	3.01	3.04
	Before	Coverage	94%	91%	91%	100%	100%	100%
		Length	1.51	.816	.432	2.93	2.90	2.87
	Delta	Coverage	94%	96%	94%	100%	100%	100%
		Length	1.46	.833	.457	2.75	2.71	2.68
500	Nonparametric	Coverage	100%	100%	100%	100%	100%	100%
		Length	4.48	3.34	2.45	2.67	2.97	2.99
	Before	Coverage	98%	93%	95%	100%	100%	100%
		Length	1.42	.735	.307	2.90	2.88	2.87
	Delta	Coverage	96%	95%	95%	100%	100%	100%
		Length	1.39	.742	.321	2.79	2.75	2.74
1000	Nonparametric	Coverage	100%	100%	100%	100%	100%	100%
		Length	4.38	3.25	2.32	2.62	2.94	2.96
	Before	Coverage	100%	96%	92%	100%	100%	100%
		Length	1.37	.701	.218	2.89	2.88	2.87
	Delta	Coverage	99%	96%	95%	100%	100%	100%
		Length	1.36	.700	.227	2.81	2.79	2.78

This table presents different measures describing the finite sample performance of confidence intervals for the coefficients of in the Normal Experiment. The label "Obs." indicates the number of observations in each sample. The number of Monte Carlo and bootstrap replications is 250.

TABLE 4. *Monte Carlo Experiments: Performance of Confidence Intervals.*

Non-Normal Experiment.

Obs.	Procedure		Covariate z ($\beta_2 = 1$)			Covariate x ($\alpha_1 = 1$)		
			.4	.2	0	.4	.2	0
250	Nonparametric	Coverage	99%	100%	100%	99%	100%	99%
		Length	4.52	3.39	2.93	7.28	3.41	3.45
	Before	Coverage	89%	87%	80%	99%	100%	100%
		Length	2.75	1.10	.891	7.05	3.21	3.13
	Delta	Coverage	98%	96%	94%	99%	100%	99%
		Length	1.86	1.06	.877	4.87	3.37	3.21
500	Nonparametric	Coverage	100%	100%	100%	97%	100%	100%
		Length	3.78	3.24	2.78	2.54	3.18	3.20
	Before	Coverage	90%	90%	87%	100%	100%	100%
		Length	1.25	.938	.734	3.27	3.07	3.03
	Delta	Coverage	97%	98%	95%	98%	100%	100%
		Length	1.25	.742	.724	3.58	2.75	3.07
1000	Nonparametric	Coverage	100%	100%	100%	97%	100%	100%
		Length	3.59	3.07	2.60	1.97	3.03	3.08
	Before	Coverage	95%	96%	95%	100%	100%	100%
		Length	1.09	.781	.579	3.11	3.01	2.97
	Delta	Coverage	98%	98%	95%	99%	100%	100%
		Length	1.08	.767	.572	3.44	3.12	2.99

This table presents different measures describing the finite sample performance of confidence intervals for the coefficients of in the Non-Normal Experiment. The label "Obs." indicates the number of observations in each sample. The number of Monte Carlo and bootstrap replications is 250.

REFERENCES

- ANGRIST, J. and A. KRUEGER (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples", *Journal of the American Statistical Association* 87, pp. 328-36.
- ARELLANO, M. and C. MEGHIR (1992): "Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets", *Review of Economic Studies*, 59(3) pp. 537-59.
- BONTEMPS, C., T. MAGNAC and E. MAURIN (2009): "Set Identified Linear Models", *IDEI Working Paper*.
- BOSTIC, R., G. STUART and G. PAINTER (2009): "Housing Wealth, Financial Wealth, and Consumption: New Evidence from Micro Data", *Regional Science and Urban Economics*, 39 (1) pp. 79-89.
- BRZOWSKI, M., M. GERVAIS, P. KLEIN and M. SUZUKI (2010): "Consumption, Income, and Wealth Inequality in Canada", *Review of Economic Dynamics*, 13 (1), pp. 52-75.
- CARROLL, C., K. DYNAN and S. KRANE (2010): "Unemployment Risk and Precautionary Wealth: Evidence from Households' Balance Sheets", *The Review of Economics and Statistics*, 85:3, pp. 586-604.
- CHEN, X., H. HONG and E. TAMER (2005): "Measurement Error with Auxiliary Data", *The Review of Economic Studies*, 72:2, pp. 343-66.
- CHEN, X., O. LINTON and I. VAN KEILEGOM (2003): "Estimation of Semiparametric Models when the Criterion Function is Not Smooth", *Econometrica*, 71:5, pp. 1591-1608.
- CHERNOZHUKOV, V., S. LEE and A. ROSEN (2009): "Intersection Bounds: Estimation and Inference", manuscript available *arxiv.org*.
- CHESHER, A. and L. NESHEIM (2006): "Review of the Literature on the Statistical Properties of Linked Datasets", *DTI Occasional Papers No 3*.
- CROSS, P. and C. MANSKI (2002): "Regressions, Short and Long", *Econometrica*, 70(2), pp. 357-68.
- DEVEREUX, P. and G. TRIPATHI (2009): "Optimally Combining Censored and Uncensored Datasets", *Journal of Econometrics*, 151(1) pp. 17-32.
- D'ORAZIO, DIZIO AND SCANU (2006); *Statistical Matching, Theory and Practice*, Wiley.
- DUGUNDJI, J. (1951); "An Extension of Tietze's Theorem", *Pacific Journal of Mathematics*, 1:3, 353-67.
- FANG, H., M. KEANE and D. SILVERMAN (2008): "Sources of Advantageous Selection: Evidence from the Medigap Insurance Market", *Journal of Political Economy*, 116(2) pp. 303-50.
- FANG, Z. and A. SANTOS (2014): "Inference on Directionally Differentiable Functions", unpublished manuscript.
- FAN, Y. and D. ZHU (2010): "Partial Identification and Confidence Sets for Functionals of the Joint Distribution of Potential Outcomes", *unpublished manuscript*.
- FAN, Y., R. SHERMAN and M. SHUM (2014): "Identifying Treatment Effects under Data Combination", *Econometrica*, 82(2), pp. 811-22.
- GOLDBERGER, A. (1991): *A Course in Econometrics*, Harvard University Press.
- HECKMAN, J., J. SMITH and N. CLEMENTS (1997): "Making the Most Out of Programme Evaluation and Social Experiments: Accounting for Heterogeneity in Program Impacts", *Review of Economic Studies*, 64(4), pp. 487-535.
- HELLERSTEIN, J. and G. IMBENS (1999): "Imposing Moment Restrictions from Auxiliary Data by Weighting", *The Review of Economics and Statistics*, 81:1, pp. 1-14.
- HIRANO, K. and J. PORTER (2012): "Impossibility Results for Non Differentiable Functionals", *Econometrica*, 80:4 pp. 1769-90.
- HIRIART-URRUTUY, J. and C. LEMARECHAL (2004): *Fundamentals of Convex Analysis*, Springer.
- HIRUWAKA, M. and A. PROKHOROV (2014): "Consistent Estimation of Linear Regression Models Using Matched Data", *Business Analytics Working Paper Series*, No WP201403.
- ICHIMURA, H. and E. MARTINEZ-SANCHIS (2010): "Identification and Estimation of GMM Models by Combining Two Data Sets", *unpublished manuscript*.
- IMBENS, G. and C. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters", *Econometrica*, 72:6, pp. 1845-1857.

- INOUE, A. and G. SOLON (2010): "Two-Sample Instrumental Variables Estimators", *The Review of Economics and Statistics*, 92:3, pp. 557-61.
- IPSOS MORI (2011): *Data Fusion - A White Paper by Ipsos MORI*, available at http://www.ipsos-mori.com/DownloadPublication/1425_IpsosMediaCT.WhitePaper.DataFusion.Jun2011.pdf
- JAPPELLI, T., J. PISHCKE and N. SOULELES (1998): "Testing for Liquidity Constraints in Euler Equations With Complementary Data Sources", *The Review of Economics and Statistics*, 80:2, pp. 251-62.
- KOMAROVA, T., D. NEKIPELOV and E. YAKOVLEV (2012): "Identification, Data Combination and the Risk of Disclosure", *unpublished manuscript*.
- MAGNAC, T. and E. MAURIN (2008): "Partial Identification in Binary Choice Models: Discrete Regressors and Interval Data", *Review of Economic Studies*, 75:4 pp. 835-64.
- MANSKI, C. (2003): *Partial Identification of Probability Distributions*, Springer.
- MEGHIR, C. and M. PALME (1999): "Assessing the Effect of Schooling on Earnings Using a Social Experiment", *The Institute for Fiscal Studies*, Working Paper No W99/10.
- MOLINARI, F. and M. PESKI (2006): "Generalization of a Result on Regressions, Short and Long", *Econometric Theory*, 22(1) pp. 159-63.
- MORIARITY, C. and F. SCHEUREN, (2003): "A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations", *Journal of Business and Economic Statistics*, 21:1, pp. 65-73.
- PACINI, D. (2016): "A Confidence Interval for a Parameter Restricted by Nonsmooth Inequalities", *manuscript*.
- POIRER, A. and N. ZIEBARTH (2014): "A Simple Estimator for Datasets with Non-Unique Identifiers", *Unpublished manuscript*.
- RACHEV, S. and L. RUSCHENDORF (1998): *Mass Transportation Problems, Volume I: Theory*, Springer.
- RASSLER, S. (2002): *Statistical Matching*, Springer.
- RIDDER, G. and R. MOFFITT (2007): "The Econometrics of Data Combination", in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Volume 6B, Elsevier.
- ROBINSON, W. (1950): "Ecological Correlation and the Behavior of Individuals", *American Sociological Review*, 15(3), pp. 351-357.
- ROCKAFELLAR, T. (1970): *Convex Analysis*, Princeton University Press.
- RODGERS, W. (1984): "An Evaluation of Statistical Matching", *Journal of Business and Economic Statistics*, 2:1, 91-102.
- RUBIN, D. (1986): "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations", *Journal of Business and Economic Statistics* 4:1, pp. 87-94 .
- SHAPIRO, (1990): "On Concepts of Directional Differentiable Functions", *Journal of Optimization Theory and Applications*, 66:3, pp. 477-87.
- THE NIELSEN COMPANY (2009): *Introduction to Nielsen Data Fusion*, available at <http://www.nielsen.com/content/dam/corporate/us/en/docs/solutions/Nielsen-Introduction-to-Data-Fusion.pdf>
- Van der VAART, A. (1998): *Asymptotic Statistics*, Cambridge University Press.
- VITALE, R. (1979): "Regression with Given Marginals", *The Annals of Statistics*, 7(3), pp. 653-658.
- WOOLDRIDGE, J. (2002): *Econometric Analysis of Cross Sections and Panel Data*, MIT Press.

APPENDIX A: PROOFS OF THE RESULTS IN SECTION 3

This appendix contains the proof of the results about identification.

Proof of Lemma 2. The support function of Θ_M is equal to:

$$s_M(q) := \sup_{\lambda \in \Lambda} q' f(\lambda) = \sup_{\lambda \in \Lambda} (v_q^o + e_q^o \lambda) = v_q^o + \sup_{\lambda \in \Lambda} e_q^o \lambda$$

Since the objective function in the programming problem in the latter display is linear, we have:

$$s_M(q) = v_q^o + e_q^o \mathbf{1}(e_q^o > 0) \sup_{\lambda \in \Lambda} \lambda + e_q^o \mathbf{1}(e_q^o \leq 0) \inf_{\lambda \in \Lambda} \lambda$$

and a solution to $\sup_{\lambda \in \Lambda} \lambda$ and $\inf_{\lambda \in \Lambda} \lambda$ must occur at the boundary of the feasible set Λ . We now characterize the boundary points of Λ .

For any random vector a and random variable b , let ρ_{ab} denote the correlation between the elements of a and b . Consider the determinant of the correlation of $(y, z', x)'$:

$$1 + \rho'_{zy} \rho_{zx} \rho_{yx} + \rho_{xy} \rho'_{zy} \rho_{zx} - \rho_{yx} \rho_{yx} - \rho'_{zx} \rho_{zx} - \rho_{zy} \rho_{zy}$$

The variance of $(y, z', x)'$ is positive semidefinite if and only if the latter determinant is nonnegative. Viewed as a function of ρ_{yx} , one can rewrite this determinant and its sign restriction as the quadratic inequality:

$$A \rho_{yx} \rho_{yx} + B \rho_{yx} + C \geq 0$$

where $A := -1$, $B := 2\rho'_{zy} \rho_{zx}$ and $C := 1 - \rho'_{zx} \rho_{zx} - \rho'_{zy} \rho_{zy}$. Since A is negative, the solution to the latter quadratic inequality is the interval defined by the two real roots of the quadratic equation $A \rho_{yx} \rho_{yx} + B \rho_{yx} + C = 0$:

$$\rho_{yx}^- := \frac{-B + \sqrt{B^2 - 4AC}}{2A}; \rho_{yx}^+ := \frac{-B - \sqrt{B^2 - 4AC}}{2A}$$

Replacing A , B and C by their definitions and rearranging terms,

$$\begin{aligned} \rho_{yx}^- &= \rho'_{zy} \rho_{zx} + \sqrt{(1 - \rho'_{zx} \rho_{zx})(1 - \rho'_{zy} \rho_{zy})} \\ \rho_{yx}^+ &= \rho'_{zy} \rho_{zx} - \sqrt{(1 - \rho'_{zx} \rho_{zx})(1 - \rho'_{zy} \rho_{zy})} \end{aligned}$$

Under Assumption P.ii, ρ_{yx}^- and ρ_{yx}^+ are finite. Plugging in the endpoints ρ_{yx}^- , ρ_{yx}^+ in $E(yx) = E(y)E(x) + [V(y)V(x)]^{1/2} \rho_{yx}$, one has that Λ is an interval with endpoints λ_{Ml}^o and λ_{Mu}^o , which are finite by Assumption (P.ii). To conclude, plug the endpoints of Λ back in the support function and notice that $e_q^o \mathbf{1}(e_q^o \leq 0) \lambda_{Ml}^o + e_q^o \mathbf{1}(e_q^o > 0) \lambda_{Mu}^o = \max(e_q^o \lambda_{Ml}^o, e_q^o \lambda_{Mu}^o)$. ■

Before proving Lemma 3, we re-state, in a notation suitable for our purposes, an existing result characterizing bounds on the expectation of the product of two random variables with given marginals.

Lemma A.1. (*Explicit Solution for the Monge-Kantorovich Problem - Rachev and Ruschendorf, 1998, Theorem 3.1.2*). Let F^o be a distribution function on \mathbb{R}^2 with marginals G_y^o and G_k^o and let (y, x_k) be distributed according to F^o . Let $\mathcal{F}_{y,k}$ denote the family of distribution func-

tions on \mathbb{R}^2 with given marginals G_y^o and G_k^o . Suppose that there is a right continuous function $y, x_k \mapsto c(y, x_k)$ satisfying the so-called Monge condition:

$$c(\tilde{y}, \tilde{x}_k) - c(y, \tilde{x}_k) - c(\tilde{y}, x_k) + c(y, x_k) \geq 0$$

for $\tilde{x}_k \geq x_k$, $\tilde{y} \geq y$, and that $E(c(y, x_k))$ exists and is finite. Then,

$$\inf_{F \in \mathcal{F}_{y,k}} \int c(y, x) dF(y, x) = \int c(y, x_k) d \max\{G_y^o(y) + G_k^o(x_k) - 1, 0\}$$

and

$$\sup_{F \in \mathcal{F}_{y,k}} \int c(y, x) dF(y, x) = \int c(y, x_k) d \min\{G_y^o(y), G_k^o(x_k)\}$$

The value function in the optimization problems above correspond to the function $F \mapsto E(c(y, x_k))$ evaluated at the Hoeffding-Frechet distributions. With Lemma A1 at hand, we proceed with the proof of Lemma 3 in the text.

Proof of Lemma 3. Using the notation in the Lemma, one can write the support function of Θ_F as:

$$s_F(q) = \sup_{F_{xy} \in \mathcal{F}} q' f \circ s(F_{xy}) = \sup_{F_{yx} \in \mathcal{F}} (v_q^o + e_q^o s(F_{yx})) = v_q^o + \sup_{F_{yx} \in \mathcal{F}} e_q^o s(F_{yx})$$

where the third equality follows because v_q^o does not depend on F_{yx} . Since e_q^o does not depend on F_{yx} ,

$$s_F(q) = v_q^o + e_q^o \inf_{F_{yx} \in \mathcal{F}} s(F_{yx}) \mathbf{1}(e_q^o \leq 0) + e_q^o \sup_{F_{yx} \in \mathcal{F}} s(F_{yx}) \mathbf{1}(e_q^o > 0)$$

The optimization problems in the latter display can be re-written as:

$$\begin{aligned} \inf_{F_{yx} \in \mathcal{F}} s(F_{yx}) &= \inf_{F_{yx|z} \in \mathcal{F}_{y|z, x|z}} \int yx dF_{y,x|z}(y, x|z) dF_z(s) ds \\ \sup_{F_{yx} \in \mathcal{F}} s(F_{yx}) &= \sup_{F_{yx|z} \in \mathcal{F}_{y|z, x|z}} \int yx dF_{y,x|z}(y, x|z) dF_z(s) ds \end{aligned}$$

Since $y, x \mapsto yx$ in the objective function satisfies the Monge Condition, it follows from Lemma A.1. (Explicit Solution to the Monge-Kantorovich Problem) that the solution occurs at the conditional Hoeffding-Frechet distributions:⁸

$$G_{yx|z}^l(y, x|s) := \max\{0, G_{y|z}^o(y|s) + G_{x|z}^o(x|s) - 1\} ; G_{yx|z}^u(y, x|s) := \min\{G_{y|z}^o(y|s), G_{x|z}^o(x|s)\}$$

The claim in the Theorem follows after evaluating $F_{yx|z} \mapsto \int yx dF_{y,z|z}(y, z|s) dG_z^o(s)$ at $G_{yx|z}^l$

⁸ $c(\tilde{y}, \tilde{x}_k) - c(y, \tilde{x}_k) - c(\tilde{y}, x_k) + c(y, x_k) = c\tilde{y}\tilde{x}_k - y\tilde{x}_k - \tilde{y}x_k + yx_k = (\tilde{y} - y)(\tilde{x}_k - x_k)$, which is non-negative whenever $\tilde{x} \geq x$ and $\tilde{y} \geq y$ as required by the Monge Condition.

and $G_{yx|z}^u$. Consider first the evaluation at $G_{yx|z}^l$:

$$\begin{aligned}\lambda_{Fl}^o &:= \int_{\mathcal{Z}} \int_{\mathcal{Y} \times \mathcal{X}} yx dG_{yx|z}^l(y, x|s) dG_z^o(s) \\ &= \int_{\mathcal{Z}} \int_{\mathcal{Y} \times \mathcal{X}} yx d \max\{0, G_{y|z}^o(y|s) + G_{x|z}^o(x|s) - 1\} dG_z^o(s)\end{aligned}$$

Let $Q_{y|z}^o(\tau, z)$ and $Q_{x|z}^o(v, z)$ denote, respectively, the conditional τ -quantile of y given z and the conditional v -quantile of x given z . By using the substitutions $y = Q_{y|z}^o(\tau, z)$ and $x = Q_{x|z}^o(v, z)$

$$\lambda_{Fl}^o = \int_{\mathcal{Z}} \int_{[0,1] \times [0,1]} Q_{y|z}^o(\tau, z) \times Q_{x|z}^o(v, z) d \max\{0, \tau + v - 1\} dG_z^o(z)$$

Since $d \max\{0, \tau + v - 1\}$ is different from zero only at $\tau + v - 1 = 0$, one has:

$$\lambda_{Fl}^o = \int_{\mathcal{Z}} \int_{[0,1]} Q_{y|z}^o(\tau, z) \times Q_{x|z}^o(1 - \tau, z) d\tau dG_z^o(z)$$

By the change-of-variable $\tau = G_{y|z}^o(y|z)$:

$$\begin{aligned}\lambda_{Fl}^o &= \int_{\mathcal{Z}} \int_{\mathcal{Y}} y \times Q_{x|z}^o(1 - G_{y|z}^o(y|z), z) dG_{yz}^o(y, z) \\ &= E[y \cdot Q_{x|z}^o(1 - G_{y|z}^o(y|z)|z)]\end{aligned}$$

where the expectation is with respect to the joint distribution of (y, z) . By a similar reasoning, the evaluation at $G_{yx|z}^l$ yields $\lambda_{Fu}^o = E[y \cdot Q_{x|z}^o(G_{y|z}^o(y|z)|z)]$.

To conclude, plug the expression for λ_{Fl}^o and λ_{Fu}^o back in the expression of the support function and note that $e_q^o \mathbf{1}(e_q^o \leq 0) \lambda_{Fl}^o + e_q^o \mathbf{1}(e_q^o > 0) \lambda_{Fu}^o = \max\{e_q^o \lambda_{Fl}^o, e_q^o \lambda_{Fu}^o\}$. ■

Proof of Proposition 1. The statement in the proposition follows after replacing $s_M(q)$ and $s_F(q)$ in Lemma 1, Equation (2), by their characterizations in Lemma 2 and Lemma 3. ■

Proof of Lemma 4. The support function of Θ_M is:

$$s_M(q) = \sup_{\lambda \in \Lambda} q' f(\lambda) = v_q^o + \sup_{\lambda \in \Lambda} e_q^o \lambda$$

The aim is to find a closed form expression for the value function in the programming problem in the latter display.

We begin by characterizing the set Λ in a way that is suitable to our purpose. Define the linear projection of y on z by $y = z'\delta + \sigma w_A$, where $\delta := E(zz')^{-1}E(zy)$, σ is the standard deviation of the residual $y - z'\delta$ and w_A is a unit variance random variable. Define the linear projection of x on z by $x = \Pi'z + \Sigma w_B$, where $\Pi := E(zz')^{-1}E(zx')$, Σ is the standard deviation of the residual $x - \Pi'z$ and w_B is a unit variance random vector of dimension $d_x \times 1$. By construction z and w_A are uncorrelated (i.e., $E(zw_A) = 0$), as well as z and w_B (i.e., $E(zw_B) = 0$). If z includes a constant, w_A and w_B have zero mean (i.e., $E(w_A) = 0$ and $E(w_B) = 0$). If the variance of (y, x', z') is positive definite so is the variance of (w_A, z', w_B') . We now write the variance of (w_A, z', w_B') in terms of the unknown expectation $\lambda = E(yx)$.

The covariance between w_A and w_B is

$$\begin{aligned}
C(w_A, w_B) &= E(w_A w_B) \\
&= E[\sigma^{-1}(y - z'\delta)\Sigma^{-1}(x - \Pi'z)] \\
&= (\sigma\Sigma)^{-1}[\lambda - \Pi'E(zy) - E(xz')\delta + \Pi'E(zz')\delta] \\
&:= A(\lambda - B) := \tilde{\lambda}
\end{aligned}$$

where the first equality follows because the residuals w_A and w_B have zero mean, and the second equality follows after replacing w_A and w_B by their definitions. Using the notation introduced above, the variance of (w_A, z', w'_B) can be written as:

$$M := \begin{pmatrix} 1 & 0_{d_z} & (\lambda - B)'A' \\ 0_{d_z} & C(z, z') & 0_{d_z \times d_x} \\ A(\lambda - B) & 0_{d_x \times d_z} & I_{d_x} \end{pmatrix}$$

where 0_{d_z} is a d_z -dimensional column vector of zeros, $0_{d_x \times d_z}$ is a $d_x \times d_z$ matrix of zeros and I_{d_x} is an $d_x \times d_x$ identity matrix. Let Ω denote the variance of (z, w_B) . If the matrix M is positive definite, so is the Schur complement of Ω in M :

$$1 - (0'_{d_z}, \tilde{\lambda}')\Omega^{-1}(0'_{d_z}, \tilde{\lambda}')'$$

Since the Schur complement of Ω in M is a scalar, the positive definite condition on it is equivalent to the inequality $1 - (0'_{d_z}, \tilde{\lambda}')\Omega^{-1}(0'_{d_z}, \tilde{\lambda}')' > 0$. For later use define the set:

$$\tilde{\Lambda} := \{\tilde{\lambda} \in R^{d_x} : (0'_{d_z}, \tilde{\lambda}')\Omega^{-1}(0'_{d_z}, \tilde{\lambda}')' \leq 1\}$$

and notice that Λ is a linear transformation of $\tilde{\Lambda}$ because $\lambda = A^{-1}\tilde{\lambda} + B$ by construction.

For a given direction q , consider now the value function $\nu(q) := \sup_{\lambda \in \Lambda} e_q \lambda$. Using the notation above,

$$\nu(q) = \sup_{\tilde{\lambda} \in \tilde{\Lambda}} e_q(A^{-1}\tilde{\lambda} + B) = e_q B + \sup_{\tilde{\lambda} \in \tilde{\Lambda}} e_q A^{-1}\tilde{\lambda}$$

where $\tilde{\lambda} := A(\lambda - B)$. Since the objective function $\tilde{\lambda} \mapsto e_q A^{-1}\tilde{\lambda}$ is linear, a solution to this programming problem has to occur at the boundary of $\tilde{\Lambda}$, that is, when $\tilde{\lambda}$ is such that

$$(0'_{d_z}, \tilde{\lambda}')\Omega^{-1}(0'_{d_z}, \tilde{\lambda}')' = 1$$

By rearranging terms, notice that the latter equality restriction is equivalent to the restriction $\|\tilde{\lambda}\| = 1$ on the norm of a vector $\tilde{\lambda}$. From the restriction $\|\tilde{\lambda}\| = 1$, we obtain then that the value of $\tilde{\lambda}$ solving the programming problem above is $\tilde{\lambda}^* = (e_q A^{-2} e'_q)^{-1/2} A^{-1} e'_q$. To conclude, replace $\tilde{\lambda} = \tilde{\lambda}^*$ in $e_q A^{-1}\tilde{\lambda}$. ■

Proof of Lemma 5. Let $s_{ky}(F_{ky})$ denote the k -element of $s(F_{1y}, \dots, F_{ky}, \dots, F_{d_x y})$. Using the

notation in the Proposition, we can write the support function as:

$$\begin{aligned}
s_F(q) &= \sup_{F_{1y}, \dots, F_{ky}, \dots, F_{d_x y} \in \mathcal{F}} q' f \circ s(F_{1y}, \dots, F_{ky}, \dots, F_{d_x y}) \\
&= v_q + \sup_{F_{1y}, \dots, F_{ky}, \dots, F_{d_x y} \in \mathcal{F}} \sum_{k=1}^{d_x} e_{q,k} s_{ky}(F_{ky}) \\
&= v_q + \sum_{k=1}^{d_x} \sup_{F_{ky} \in \mathcal{F}} e_{q,k} s_{ky}(F_{ky})
\end{aligned}$$

where the third equality follows from linearity of the function $F_{1y}, \dots, F_{d_x y} \mapsto s(F_{1y}, \dots, F_{d_x y})$. The rest of the proof parallels that of Lemma 2 and is not repeated here. ■

Proof of Theorem 1. The statement in the theorem follows after replacing the support functions in Lemma 1 by their characterizations derived in Lemma 4 and Lemma 5. ■

APPENDIX B: PROOFS OF THE RESULTS IN SECTION 4

We begin by calculating the directional differential of the bounding functions using the following chain rule.

Lemma B.1 (*Chain Rule for Hadamard Directional Differentiable Functions - Shapiro, 1990, Proposition 3.6*). *Let $h \mapsto \phi(h)$ be Hadamard directional differentiable at h_o , and let $\phi \mapsto \Phi(\phi)$ be Hadamard directional differentiable at $\phi_o := \phi(h_o)$. Let $\dot{\phi}(h_o, d)$ and $\dot{\Phi}(\phi_o, r)$ denote the Hadamard directional differential of $h \mapsto \phi(h)$ and $\phi \mapsto \Phi(\phi)$, respectively. Then, the composite mapping $h \mapsto g(h) := \Phi \circ \phi(h)$ is Hadamard directional differentiable at h_o and the chain rule $\dot{g}(h_o, d) = \dot{\Phi}(\phi_o, \dot{\phi}(h_o, d))$ holds.*

To employ the chain rule, set $h_o = \eta_{l_o}$ and $g(h) = m_l(\eta_l)$. Re-write $\eta_l \mapsto m_l(\eta_l)$ as the composition of two functions. The first function is

$$\eta_l \mapsto \phi(\eta_l) := \begin{pmatrix} \phi_M(\eta_{q_l}) \\ \phi_F(\eta_{q_l}) \end{pmatrix} := \begin{pmatrix} v_{q_l} + e_{q_l} B + (e_{q_l} A^{-2} e'_{q_l})^{1/2} \\ \sum_{k=1}^{d_x} \max_{r \in \{l, u\}} v_{q_l} + e_{q_l} \lambda_{Fkr} \end{pmatrix}$$

For a given vector $\phi := (\phi_M, \phi_F) \in \mathbb{R}^2$, the second function is

$$\phi \mapsto \Phi(\phi) := -\min\{\phi_M, \phi_F\}$$

With this notation at hand, $m_l(\eta_l) = \Phi \circ \phi(\eta_l)$. The Hadamard directional differential of $\phi \mapsto \Phi(\phi)$ at $\phi_o := (\phi_M^o, \phi_F^o)$ in the direction $r := (r_M, r_F)$ is:

$$\dot{\Phi}(\phi_o, r) := \begin{cases} -r_{t^*} & \text{if } \phi_M^o \neq \phi_F^o \\ -\min\{r_M, r_F\} & \text{if } \phi_M^o = \phi_F^o \end{cases}$$

with $t^* := \arg \min_{t \in \{M, F\}} \phi_t^o$. For ι denoting a conformable vector of ones, the Hadamard directional differential of $\eta_l \mapsto \phi_M(\eta_l)$ at η_l^o in the direction d is:

$$\dot{\phi}_M(\eta_l^o, d) = \iota' d$$

Using the chain rule above, the Hadamard directional differential of $\eta_l \mapsto \phi_F(\eta_l)$ at η_l^o in the direction d is:

$$\dot{\phi}_F(\eta_l^o, d) = \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d)$$

with

$$\dot{\phi}_{Fk}(\eta_l^o, d) := \begin{cases} d_1 + d_{ks_k^*} & \text{if } e_{q,k} \lambda_{Fkl} \neq e_{q,k} \lambda_{Fku} \\ d_1 + \max(d_{kl}, d_{ku}) & \text{if } e_{q,k} \lambda_{Fkl} = e_{q,k} \lambda_{Fku} \end{cases}$$

where d_1 , d_{kl} and d_{ku} are the 1st, $2+k$ -th and $2+k+1$ -th elements in d , respectively, and $s_k^* := \arg \max_{t \in \{l, u\}} e_{q,k} \lambda_{Fkt}$. The chain rule in Lemma B.1 then implies that the Hadamard directional differential of $\eta_l \mapsto m_l(\eta_l)$ at η_l^o in the direction d is:

$$\dot{m}_l(\eta_l^o, d) = \dot{\Phi}\left(\phi_o, r = (\dot{\phi}_M(\eta_l^o, d), \dot{\phi}_F(\eta_l^o, d))\right)$$

Using a similar reasoning, one can obtain the Hadamard directional differential of $\eta_u \mapsto m_u(\eta_u)$. We write a more explicit expression in the following Lemma:

Lemma B.2 (Hadamard Directional Differential). The Hadamard directional differential of $\eta_l \mapsto m_l(\eta_l)$ and of $\eta_u \mapsto m_u(\eta_u)$ are

$$\dot{m}_l(\eta_l^o, d) := \begin{cases} -\iota' d & \text{if } s_M(\eta_l^o) - s_F(\eta_l^o) < 0 \\ -\sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d) & \text{if } s_M(\eta_l^o) - s_F(\eta_l^o) > 0 \\ -\min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_l^o, d)\right) & \text{if } s_M(\eta_l^o) - s_F(\eta_l^o) = 0 \end{cases}$$

and

$$\dot{m}_u(\eta_u^o, d) := \begin{cases} \iota' d & \text{if } s_M(\eta_u^o) - s_F(\eta_u^o) < 0 \\ \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_u^o, d) & \text{if } s_M(\eta_u^o) - s_F(\eta_u^o) > 0 \\ \min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\eta_u^o, d)\right) & \text{if } s_M(\eta_u^o) - s_F(\eta_u^o) = 0 \end{cases}$$

respectively

The Hadamard directional differential of a Lipschitz continuous function, like $d \mapsto \dot{m}_b(\eta_b^o, d)$, is so-called the **Bouligand directional differential** (see Shapiro, 1990).

For a positive sequence δ_n diverging to infinity and δ_n/\sqrt{n} converging to zero, we estimate the Hadamard directional differentials by:

$$\hat{\dot{m}}_l(\hat{\eta}_l, d) := \begin{cases} -\iota' d & \text{if } s_M(\hat{\eta}_l) - s_F(\hat{\eta}_l) < -\delta_n \\ -\sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_l, d) & \text{if } s_M(\hat{\eta}_l) - s_F(\hat{\eta}_l) > \delta_n \\ -\min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_l, d)\right) & \text{if } -\delta_n < s_M(\hat{\eta}_l) - s_F(\hat{\eta}_l) < \delta_n \end{cases}$$

and

$$\hat{\dot{m}}_u(\hat{\eta}_u, d) := \begin{cases} \iota' d & \text{if } s_M(\hat{\eta}_u) - s_F(\hat{\eta}_u) < -\delta_n \\ \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_u, d) & \text{if } s_M(\hat{\eta}_u) - s_F(\hat{\eta}_u) > \delta_n \\ \min\left(\iota' d, \sum_{k=1}^{d_x} \dot{\phi}_{Fk}(\hat{\eta}_u, d)\right) & \text{if } -\delta_n < s_M(\hat{\eta}_u) - s_F(\hat{\eta}_u) < \delta_n \end{cases}$$

The next proposition establishes the consistency of $\hat{\dot{m}}_l$ and $\hat{\dot{m}}_u$ uniformly over d . To avoid measurability issues, we use P to denote either the probability measure of the data or the corresponding outer probability measure.

Lemma B.3 (Uniform Consistent Estimator of the Directional Differential). Let Assumptions P and D hold. Let further assume that $\delta_n \uparrow \infty$, $n^{-1/2}\delta_n \downarrow 0$, and the conditions C.1-C.4 in Theorem 2 (Asymptotic Properties of the Nuisance Parameter Estimator) hold. Then,

$$\|\hat{\dot{m}}(\hat{\eta}_n, d) - \dot{m}(\eta_o, d)\| = o_{P_o}(1)$$

for all $d \in \mathbb{R}^{10}$.

Proof. Under conditions C.1-C.4, by the continuous mapping theorem, $s_t(\hat{\eta}_b)$ converges in probability to $s_t(\eta_b^o)$ because $\hat{\eta}_b$ converges in probability to η_b^o and $\eta_b \mapsto s_t(\eta_b)$ is continuous for $b \in t, l$ and $t \in M, F$. Under $\delta_n \uparrow \infty$ and $n^{-1/2}\delta_n \downarrow 0$ then,

$$\|\hat{\dot{m}}_b(\hat{\eta}_n, d) - \dot{m}_b(\eta_o, d)\| = o_{P_o}(1)$$

Since the result in the latter display holds jointly for $b \in \{l, u\}$, they can be combined to obtain the claim in Lemma B.3. ■

To establish theoretical properties of the bias-corrected estimator, we verify the conditions of a result in Fang and Santos (2014). For the sake of completeness, we begin by re-stating this result in a notation suitable for our purposes:

Lemma B.4 (Consistency of the Delta Bootstrap - Fang and Santos, 2014, Theorem 3.3). Consider a function $f : \mathbb{D}_g \subseteq \mathbb{D} \mapsto \mathbb{E}$ describing a parameter of interest $f_o := f(h_o)$, where \mathbb{D}_f denotes the domain of f and h_o a nuisance parameter. Let assume that:

FS Condition 2.1.

(i) \mathbb{D} and \mathbb{E} are Banach spaces with norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$, respectively.

(ii) $f : \mathbb{D}_f \subseteq \mathbb{D} \mapsto \mathbb{E}$ is Hadamard directional differentiable at h_o tangentially to the set $\mathbb{D}_o \in \mathbb{D}$, where \mathbb{D}_o is a subset of \mathbb{D} .

FS Condition 2.2.

(i) $h_o \in \mathbb{D}_f$ and there is an estimator \hat{h}_n of this nuisance parameter such that, for some sequence of positive numbers $r_n \uparrow \infty$, the random element $r_n(\hat{h}_n - h_o)$ converges in distribution to a random element \mathbb{G}_o in \mathbb{D} .

(ii) \mathbb{G}_o is tight and its support is included in \mathbb{D}_o .

FS Condition 2.3.

(i) The Hadamard directional derivative $\dot{f}(h_o, d)$ can be continuously extended to \mathbb{D} .

FS Condition 3.1. Let \hat{h}_n^* denote the bootstrapped version of the estimator \hat{h}_n . Let $\{X_l\}_{l=1}^n$ denote the data and let $\{W_l\}_{l=1}^n$ denote random weights.

(i) $\hat{h}_n^* : \{X_l, W_l\}_{l=1}^n \mapsto \mathbb{D}_\theta$ with $\{X_l\}_{l=1}^n$ independent of $\{W_l\}_{l=1}^n$.

(ii) \hat{h}_n^* satisfies

$$\sup_{b \in BL_1(\mathbb{D})} |E[b(r_n(\hat{h}_n^* - \hat{h}_n)) | \{X_l\}_{l=1}^n] - E[f(\mathbb{G}_o)]| = o_P(1),$$

where $BL_1(\mathbb{D})$ is the set of Lipschitz functionals from \mathbb{D} to \mathbb{R} whose level and Lipschitz constant are bounded by one.

FS Condition 3.2.

(i) The sequence $r_n(\hat{h}_n^* - \hat{h}_n)$ is asymptotically measurable.

(ii) $b(r_n(\hat{h}_n^* - \hat{h}_n))$ is a measurable function of $\{W_l\}_{l=1}^n$ outer almost surely in $\{X_l\}_{l=1}^n$ for any continuous and bounded function $b : \mathbb{D} \mapsto \mathbb{R}$.

FS Condition 3.3. There is an estimator \hat{f} of \dot{f} such that for every compact set $K \subseteq \mathbb{D}_o$, $K^\delta := \{a \in \mathbb{D} : \inf_{b \in K} \|a - b\|_{\mathbb{D}} < \delta\}$, and every $\epsilon > 0$:

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} P_o \left(\sup_{d \in K^\delta} \left\| \hat{f}(\hat{h}_n, d) - \dot{f}(h_o, d) \right\|_{\mathbb{E}} > \epsilon \right) = 0$$

Then,

$$\sup_{b \in BL_1(\mathbb{E})} \left| E \left[b \left(\hat{f}(\hat{h}_n, r_n(\hat{h}_n^* - \hat{h}_n)) \right) \mid \{X_l\}_{l=1}^n \right] - E \left[b(\dot{f}(h_o, \mathbb{G}_o)) \right] \right| = o_P(1)$$

To prove Theorem 2, we now verify the conditions of Lemma B.4.

Proof of Theorem 2. Our case corresponds to $h_o = \eta_b^o$ and $f(h) = m_b(\eta_b)$ with $\eta_b \in \mathbb{D} = \mathbb{R}^{2+2d_x}$, $\mathbb{E} = \mathbb{R}$, and $\mathbb{D}_g = \mathbb{K}^{2+2d_x}$ for some set compact set $\mathbb{K}^{2+2d_x} \subseteq \mathbb{R}^{2+2d_x}$.

To verify FS Condition 2.1(i), let equip \mathbb{D} and \mathbb{E} with the sup-norm. Then, \mathbb{D} and \mathbb{E} are complete normed vector spaces (i.e., Banach spaces; see Rudin, 1986, Chapter 5.1). To verify FS Condition 2.1(ii), we refer to our discussion of Lemma B.2 (Hadamard Directional Differential).

To verify FS Condition 2.2, we first notice that, by construction, the nuisance parameter $h_o = \eta_b^o$ lives in some compact set. To verify FS Condition 2.2 (i), without loss of generality one can set \mathbb{K}^{2+2d_x} equal to that compact set. The estimator is $\hat{h}_n = \hat{\eta}_b$. Under the conditions of Lemma C.3 (Asymptotic Properties of the Nuisance Parameter Estimator), we can set $r_n = n_A^{1/2}$ and \mathbb{G}_o equal to a multivariate normal random vector. To verify FS Condition 2.2 (ii), it suffices to note that a multivariate normal random vector \mathbb{G}_o is tight (because any random vector \mathbb{G}_o is tight: for every constant $\epsilon > 0$ there exists a constant κ such that $P_o(\|\mathbb{G}_o\| > \kappa) < \epsilon$) and its support belongs to \mathbb{D}_o .

To verify FS Condition 2.3.(i), note that \mathbb{D}_o corresponds to a cone in an Euclidean space. The cone \mathbb{D}_o is a closed set in the topology induced by the sup norm $\|\cdot\|_{\mathbb{D}}$. Since $\eta_b \mapsto m_b(\eta_b)$ is Lipchitz continuous, $d \mapsto \dot{m}_b(\eta_b^o, d)$ is continuous (see e.g., Shapiro, 1990). Since \mathbb{D}_o is closed, then the continuity of $d \mapsto \dot{m}_b(\eta_b^o, d)$ and Theorem 4.1 in Dugundji (1951) imply that \dot{m}_b admits a continuous extension to \mathbb{D} .

To verify FS Condition 3.1 (i), we refer to our discussion at the end of Appendix C about the validity of the non-parametric bootstrap to approximate the sampling distribution of the estimator of the nuisance parameters $\hat{\eta}_b$. To verify FS Condition 3.1 (ii), it suffices to prove the consistency of the law of $n_A^{1/2}(\hat{\eta}_b^* - \hat{\eta}_b)$ conditional on the data for the distribution of \mathbb{G}_o . This consistency result is established in Lemma C.4 (Consistency of the Nonparametric Bootstrap).

In our case FS Condition 3.2 is satisfied by construction because $\hat{h}_n = \hat{\eta}_b$ and $\hat{h}_n^* = \hat{\eta}_b^*$ correspond to empirical and bootstrapped empirical processes, respectively.

To verify FS Condition 3.3.(i), notice that $\|\dot{m}_b(\hat{\eta}_b, d_1) - \dot{m}_b(\hat{\eta}_b, d_2)\| \leq \kappa \|d_1 - d_2\|$ for some constant $\kappa > 0$ and all $d_1, d_2 \in \mathbb{D}$ (because $d \mapsto \dot{m}_b(\hat{\eta}_b, d)$ is Lipchitz continuous). By a result in Fang and Santos (2014, Lemma A.6), showing that, for any $d \in \mathbb{D}_o$, $\|\dot{m}_b(\hat{\eta}_b, d) - \dot{m}_b(\eta_b^o, d)\|_{\mathbb{E}} = o_{P_o}(1)$ suffices for establishing condition 3.3.(i). This latter condition is verified in Lemma B.3 (Uniform Consistent Estimator of the Directional Differential).

Hence, Lemma B.4 (Consistency of the Delta Bootstrap) implies that:

$$\begin{aligned} S^{-1} \sum_{s=1}^S \hat{\xi}_{bs}^* &= E\left(\dot{m}_b(\hat{\eta}_b, n_A^{1/2}(\hat{\eta}^* - \hat{\eta}_n))\right) + o_{P^*}(1) \\ &= E\left(\dot{m}_b(\eta_b^o, n_A^{1/2}(\hat{\eta}^* - \hat{\eta}_n))\right) + o_{P^*}(1) + o_{P_o}(1) \\ &= n_A^{1/2} E(m_b(\hat{\eta}_b) - m_b(\eta_b^o)) + o_{P^*}(1) + o_{P_o}(1) + o_{P_o}(1) \end{aligned}$$

where the second equality follows because \hat{m}_b is a consistent estimator of m_b (see Lemma B.3 Uniform Consistent Estimator of Directional Differential), and the last equality follows from the Delta Method Theorem in Fang and Santos (2014, Theorem 2.1). ■

To establish theoretical properties of the confidence interval C_n , we verify the conditions of a result in a companion paper (Pacini, 2016). For the sake of completeness, we begin by re-stating this result in a notation suitable for our purposes:

Lemma B.5 (*Locally Uniform Confidence Interval - Pacini, 2016, Theorem 3*). *Let $\eta_o := h(P_o)$ be an unknown nuisance parameter defined by a known bijective function $h : \mathcal{P} \mapsto \mathcal{H}$ taking*

values in a space \mathcal{H} . The unknown parameter of interest $\theta_o \in \Theta \subset \mathbb{R}$ satisfies the inequalities

$$m_l(\eta_o) =: \theta_l \leq \theta_o \leq \theta_u := m_u(\eta_o)$$

where $m_b : \mathcal{H} \mapsto \mathbb{R}$ for $b \in \{l, u\}$ are known (up to η_o) bounding functions. Let the following assumptions hold:

Pa Condition 2. For each $b \in \{l, u\}$, for some Lipchitz constant $L_b > 0$ and any η_1 and $\eta_2 \in \mathcal{H}$, $\eta \mapsto m_b(\eta)$ is L -continuous:

$$\|m_b(\eta_1) - m_b(\eta_2)\|_{\mathbb{R}} \leq L_b \|\eta_1 - \eta_2\|_{\mathcal{H}} \quad (\text{Pa 2.i})$$

and Hadamard directional differentiable at η_o , in the sense that, the limit

$$\dot{m}_b(\eta_o, d) := \lim_{t_n \downarrow 0} t_n^{-1} [m_b(\eta_o + t_n d) - m_b(\eta_o)] \quad (\text{Pa 2.ii})$$

is finite for all $d \in \mathcal{H}$, with \mathcal{H} a complete normed vector (Banach) space.

Pa Condition 3. There is an estimator $\hat{m}_b(\hat{\eta}_n, d)$ of $\dot{m}_b(\eta_o, d)$ satisfying:

$$\sup_{d \in K^\delta} |\hat{m}_b(\hat{\eta}_n, d) - \dot{m}_b(\eta_o, d)| = o_{P_o}(1) \quad \text{for any } \delta\text{-enlargement } K^\delta \text{ of a compact set } K \in \mathcal{H} \quad (\text{Pa 3})$$

Pa Condition 4. There is an estimator $\hat{\eta}_n : \{w_i\}_{i=1}^n \mapsto \mathcal{H}$ of η_o . For $P_n \in \mathcal{P}_{n,h}$, $\eta_n = h(P_n)$ and Z_{η_o} denoting a tight random element taking values in \mathcal{H} , $\hat{\eta}_n$ satisfies:

$$Z_{P_n, n} := n^{1/2}(\hat{\eta}_n - \eta_n) \underset{P_n}{\rightsquigarrow} Z_{\eta_o} \quad (\text{Pa 4})$$

where the distribution of Z_{η_o} can depend on η_o but not on η_n .

Pa Condition 5. There is an approximation \hat{Z}_n^* of Z_{η_o} satisfying

$$\sup_{b \in BL_{\mathcal{H}}} E[b(\hat{Z}_n^*) | \{w_i\}_{i=1}^n] - E[b(Z_{\eta_o})] = o_{P_o}(1) \quad (\text{Pa 5})$$

Consider the confidence interval C_n constructed according to:

Step 1. For a large S , simulate $1, \dots, s, \dots, S$ realizations of \hat{Z}_n^* . Denote a given realization by \hat{Z}_{ns}^* .

Step 2. Calculate $\hat{T}_{ns}^* := \max(\hat{m}_l(\hat{\eta}_n, \hat{Z}_{ns}^*), 0)^2 + \min(\hat{m}_u(\hat{\eta}_n, \hat{Z}_{ns}^*), 0)^2$.

Step 3. Fix $\tau \in (0, 1)$. Set $\hat{q}_{1-\alpha}$ equal to the $1 - \alpha$ empirical quantile of $\{\hat{T}_{ns}^*\}_{s=1}^S$.

Step 4. Create a grid in Θ . Let θ_c denote a point in this grid.

Step 5. Calculate $T_n(\theta_c)$ for each θ_c in the grid. Accept θ_c if $T_n(\theta_c) < \hat{q}_{1-\alpha}$ otherwise discard θ_c . Take the smallest and largest accepted values as the endpoints of C_n .

If

Pa Condition 6. The limiting distribution of the test statistic $T_n(\theta)$ is continuous and strictly increasing at its $1 - \alpha$ quantile $q_{1-\alpha}$.

Pa Condition 7. For some $(\eta_o, \theta_o) \in \mathcal{H}_n \times \Theta_o$, $\xi_o = (0, 0)$.

Then, C_n satisfies

$$\liminf_{n \rightarrow \infty} \inf_{P \in \mathcal{P}_{n,h}, \theta \in [\theta_l, \theta_u]} P(\theta \in C_n) = 1 - \tau$$

Proof of Theorem 3. For $\text{sum}_{k=1}^{d_x} := \sum_{k=1}^{d_x}$, define $g_{1l}(\eta_1, \eta_2) := \text{sum}_{k=1}^{d_x} \max(\eta_{1k}, \eta_{2k})$, $r_{1l}(g_{1l}) := -g_{1l}$, $g_{2l}(r_{1l}, \eta_3) := \max(r_{1l}, -\eta_3)$, $r_{2l}(g_{2l}) := -g_{2l}$. Rewrite the lower bounding

function as

$$m_l(\eta) = -\min\left(\sum_{k=1}^{d_x} \max(\eta_{1k}, \eta_{2k}), \eta_3\right) = r_{2l}\left(g_{2l}\left(r_{1l}(g_{1l}(\eta_1, \eta_2), \eta_3)\right)\right)$$

We have already verified that η belongs to a Banach space (See Proof of Theorem 2). Since the *max* and *sum* functions are L-continuous, and L-continuity is preserved under composition of L-continuous functions, $\eta_1, \eta_2 \mapsto g_{1l}(\eta_1, \eta_2)$ and $r_{1l}, \eta_3 \mapsto g_{2l}(r_{1l}, \eta_3)$ are L-continuous (Pa 2.i). To verify Pa 2.ii, we refer to our discussion of Lemma B.2 (Hadamard Directional Differential) To verify Pa 3, we refer to Lemma B.3 (Uniform Consistent Estimator of the Directional Differential) and the verification of FS Condition 3.3 in the proof of Theorem 2. The nuisance parameters η_o are population moments and their sample analogs satisfy Pa 4 with Z_{η_o} a zero-mean multivariate normal vector with asymptotic variance depending on η_o (see Lemma C.3 Asymptotic Properties of Nuisance Parameter Estimator). The approximation \hat{Z}_n^* in Pa 5 can be constructed using the bootstrap (see Lemma C.4 Consistency of the Nonparametric Bootstrap). Since in this application \hat{Z}_{η_o} is Gaussian (see Lemma C.3 Asymptotic Properties of Nuisance Parameter Estimator) and $\dot{m}_b(\eta_b^o)$ is non zero, Pa 3 and Pa 4 by a result in Davydov, Lifschitz and Smorodina (1998, Theorem 11.1) imply Pa 6. Pa 7 is satisfied when either the covariates observed in different samples are uncorrelated or there is an instrumental variable observed in both samples (see the discussion of point identification in Section 3). ■

APPENDIX C: ESTIMATION OF NUISANCE PARAMETERS

This appendix describes estimators for the nuisance parameters and establishes some of its asymptotic properties. These asymptotic properties in turn are useful to establish the properties of the inference procedures described in the text.

We begin by describing the estimators of components of the support functions. The estimators are:

$$\begin{aligned}
\hat{d} &:= -(\hat{s}_{zz'} - \hat{s}_{zx'}\hat{s}_{xx'}^{-1}\hat{s}_{xz'})^{-1}\hat{s}_{zx'}\hat{s}_{xx'}^{-1} \\
\hat{c} &:= [\hat{s}_{zz'} - \hat{s}_{zx'}\hat{s}_{xx'}^{-1}\hat{s}_{xz'}]^{-1}\hat{s}_{zy} \\
\hat{b} &:= -\hat{s}_{xx'}^{-1}\hat{s}_{xz'}\hat{d} - \hat{s}_{xx'}^{-1}\hat{s}_{xz'}\hat{c} \\
\hat{a} &:= \hat{s}_{xx'}^{-1} - \hat{s}_{xx'}^{-1}\hat{s}_{xz'}\hat{d} \\
\hat{e}_q &:= q'_\alpha\hat{b} + q'_\beta\hat{d} \\
\hat{v}_q &:= q'_\alpha\hat{a} + q'_\beta\hat{c} \\
\hat{A} &:= (\hat{\sigma}\hat{\Sigma})^{-1} \\
\hat{B} &:= \hat{\Pi}\hat{s}_{zy} - \hat{s}_{zx'}\hat{\delta} + \hat{\Pi}\hat{s}_{zz'}\hat{\delta} \\
\hat{\lambda}_{Fkl} &:= n_A^{-1} \sum_{i=1}^{n_A} y_i \hat{Q}_k (1 - \hat{G}(y_i|z_i)|z_i) \\
\hat{\lambda}_{Fku} &:= n_A^{-1} \sum_{i=1}^{n_A} y_i \hat{Q}_k (\hat{G}(y_i|z_i)|z_i),
\end{aligned}$$

where $\hat{s}_{zz'} := n^{-1} \sum_{i=1}^n z_i z'_i$, $\hat{s}_{zx'} := n_B^{-1} \sum_{j=n_A+1}^n z_j x'_j$, $\hat{s}_{xx'} := n_B^{-1} \sum_{j=n_A+1}^n x_j x'_j$, $\hat{s}_{yz} := n_A^{-1} \sum_{i=1}^{n_A} y_i z_i$ and $y, z \mapsto \hat{G}(y|z)$ and $\tau, z \mapsto \hat{Q}_k(\tau|z)$ are non-parametric estimators for the conditional distribution $y, z \mapsto G_{y|z}^o(y|z)$ and the conditional quantile function $\tau, z \mapsto Q_{k|z}^o(\tau|z)$, respectively.

Let define $\hat{\eta} := (\hat{\eta}'_{q_l}, \hat{\eta}'_{q_u})'$ with

$$\begin{aligned}
\hat{\eta}_{q_l} &:= (\hat{v}_{q_l}, \hat{e}_{q_l}\hat{B} + (\hat{e}_{q_l}\hat{A}^{-2}\hat{e}_{q_l})^{1/2}, \hat{e}_{q_l,1}\hat{\lambda}_{F1l}, \hat{e}_{q_l,1}\hat{\lambda}_{F1u}, \dots, \hat{e}_{q_l,d_x}\hat{\lambda}_{Fd_xl}, \hat{e}_{q_l,d_x}\hat{\lambda}_{Fd_xu})' \\
\hat{\eta}_{q_u} &:= (\hat{v}_{q_u}, \hat{e}_{q_u}\hat{B} + (\hat{e}_{q_u}\hat{A}^{-2}\hat{e}_{q_u})^{1/2}, \hat{e}_{q_u,1}\hat{\lambda}_{F1l}, \hat{e}_{q_u}\hat{\lambda}_{F1u}, \dots, \hat{e}_{q_u,1}\hat{\lambda}_{Fd_xl}, \hat{e}_{q_u,1}\hat{\lambda}_{Fd_xu})'
\end{aligned}$$

The components of $\hat{\eta}$ are either sample analogs of unconditional moments (e.g., \hat{v}_{q_l}) or sample analogs of unconditional moments with unknown functions estimated non-parametrically (e.g., $\hat{e}_{q_l,k}\hat{\lambda}_{Fkl}$). If there were no unknown functions estimated nonparametrically, convergence in probability for $\hat{\eta}$ and convergence in distribution for $(\hat{\eta} - \eta_o)\sqrt{n_A}$ could be established using a Law of Large Numbers and a Central Limit Theorem for sums of i.i.d. random variables. The presence of unknown functions estimated nonparametrically makes the direct application of these standard results an inviable approach. A viable approach is to use the Law of Large Numbers and the Central Limit Theorem for semiparametric estimation problems that involve a criterion function depending on both finite dimensional and infinite dimensional unknown parameters as in Chen, Linton and Van Keilegom (2003, Theorems 1 and 2). We next re-state these results in a notation suitable for our purposes.

Lemma C.1 (Law of Large Numbers - Chen, Linton and Van Keilegom, 2003, Theorem 1). Assume that data $\{X_l\}_{l=1}^n$ is randomly sampled from a distribution P whose support is a proper subset of \mathbb{R}^d . Let denote \mathcal{A} for a finite dimensional parameter set and Ψ for an infinite dimensional parameter set. Let equip these sets with norms $\|\cdot\|_{\mathcal{A}}$ and $\|\cdot\|_{\Psi}$, re-

spectively. For some positive integers p and k , assume that there exists a measurable vector-valued function $X_l, \gamma, \psi \mapsto m(X_l, \gamma, \psi) : \mathbb{R}^d \times \mathbb{R}^p \times \Psi \mapsto \mathbb{R}^k$, and define the nonrandom function $\gamma, \psi \mapsto M(\gamma, \psi) := E(m(X_l, \gamma, \psi)) : \mathbb{R}^p \times \Psi \mapsto \mathbb{R}^k$. Let denote $\gamma_o \in \mathcal{A}$ and $\psi_o \in \Psi$ as the true unknown finite and infinite dimensional parameters. Define the sample analog $M_n(\gamma, \psi) := n^{-1} \sum_{l=1}^n m(X_l, \gamma, \psi)$ and assume there is a nonparametric estimator $\hat{\psi}$ of ψ_o . Define the estimator $\hat{\gamma} := \arg \min_{\gamma} \|M_n(\gamma, \hat{\psi})\|$. Suppose further that:

CLK Condition (1.1). $\|M_n(\hat{\gamma}, \hat{\psi})\| \leq \inf_{\gamma \in \mathcal{A}} \|M_n(\gamma, \hat{\psi})\| + o_P(1)$.

CLK Condition (1.2). For all positive constant $\delta > 0$, there exists another positive constant $\epsilon(\delta) > 0$ such that $\inf_{\|\gamma - \gamma_o\|_{\mathcal{A}} > \delta} \|M(\gamma, \psi_o)\| \geq \epsilon(\delta) > 0$.

CLK Condition (1.3). Uniformly for all $\gamma \in \mathcal{A}$, $\psi \mapsto M(\gamma, \psi)$ is continuous (with respect to the metric $\|\cdot\|_{\Psi}$) at $\psi = \psi_o$.

CLK Condition (1.4). $\|\hat{\psi} - \psi_o\|_{\Psi} = o_P(1)$.

CLK Condition (1.5'). For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,

$$\sup_{\gamma \in \mathcal{A}, \|\psi - \psi_o\|_{\Psi} < \delta_n} \|M_n(\gamma, \psi) - M(\gamma, \psi)\| = o_P(1)$$

Then, $\hat{\gamma} - \gamma_o = o_P(1)$.

Lemma C.2 (Central Limit Theorem - Chen, Linton and Van Keilegom, 2003, Theorem 2). Assume that the conditions of the Law of Large Numbers in Lemma C.1 are satisfied. For some sequence $\{\delta_n\}$ of positive numbers converging to zero, define the shrinking sets $\mathcal{A}_{\delta_n} := \{\gamma \in \mathcal{A} : \|\gamma - \gamma_o\|_{\mathcal{A}} \leq \delta_n\}$ and $\Psi_{\delta_n} := \{\psi \in \Psi : \|\psi - \psi_o\|_{\Psi} \leq \delta_n\}$. For any (γ, ψ) , let denote the derivative of $\gamma \mapsto M(\gamma, \psi)$ evaluated at γ by $\Gamma_1(\gamma, \psi)$ and the pathwise derivative of $\psi \mapsto M(\gamma, \psi)$ at ψ in the direction $(\bar{\psi} - \psi)$ by $\Gamma_2(\gamma, \psi)[\bar{\psi} - \psi]$. Assume further that:

CLK Condition (2.1). $\|M_n(\hat{\gamma}, \hat{\psi})\| \leq \inf_{\gamma \in \mathcal{A}} \|M_n(\gamma, \hat{\psi})\| + o_P(n^{-1/2})$.

CLK Condition (2.2). (i) The derivative $\gamma \mapsto \Gamma_1(\gamma, \psi_o)$ exists for $\gamma \in \mathcal{A}_{\delta_n}$ and is continuous at $\gamma = \gamma_o$;

(ii) The matrix $\Gamma_1 := \Gamma_1(\gamma_o, \psi_o)$ is of full (column) rank.

CLK Condition (2.3). For all $\gamma \in \mathcal{A}_{\delta_n}$, the pathwise derivative $\Gamma_2(\gamma, \psi_o)[\psi - \psi_o]$ of $M(\gamma, \psi_o)$ exists in all directions $[\psi - \psi_o] \in \Psi$; and for all $(\gamma, \psi) \in \mathcal{A}_{\delta_n} \times \Psi_{\delta_n}$ with a sequence δ_n of positive numbers converging to zero:

(i) $\|M(\gamma, \psi) - M(\gamma, \psi_o) - \Gamma_2(\gamma, \psi_o)[\psi - \psi_o]\| \leq \kappa \|\psi - \psi_o\|_{\Psi}^2$ for some positive constant $\kappa > 0$;

(ii) $\|\Gamma_2(\gamma, \psi_o)[\psi - \psi_o] - \Gamma_2(\gamma_o, \psi_o)[\psi - \psi_o]\| \leq o(1)\delta_n$.

CLK Condition 2.4. $\hat{\psi} \in \Psi$ with probability tending one; and $\|\hat{\psi} - \psi_o\|_{\Psi} = o_P(n^{-1/4})$.

CLK Condition 2.5'. For all sequences $\{\delta_n\}$ of positive numbers converging to zero

$$\sup_{\|\gamma - \gamma_o\|_{\mathcal{A}} < \delta_n, \|\psi - \psi_o\|_{\Psi} < \delta_n} \|M_n(\gamma, \psi) - M(\gamma, \psi) - M_n(\gamma_o, \psi_o)\| = o_P(n^{-1/2})$$

CLK Condition 2.6'. $\sqrt{n}(M_n(\gamma_o, \psi_o) + \Gamma_2(\gamma_o, \psi_o)[\hat{\psi} - \psi_o])$ converges in distribution to a random vector with multivariate normal distribution.

Then, $(\hat{\gamma} - \gamma_o)\sqrt{n}$ converges in distribution to a random vector with multivariate normal distribution.

We next verify the conditions of these two Lemmas to obtain the following asymptotic properties for $\hat{\eta}$:

Lemma C.3 (Asymptotic Properties of the Nuisance Parameter Estimator $\hat{\eta}$). Let Assumptions P and D hold. Set the infinite dimensional parameter space Ψ equal to the product of the spaces

of functions ψ_k mapping the support of (y, z) into the support of (x_k, x_k) according to:

$$\psi_k(y, z) := \begin{pmatrix} \psi_{1k}(y, z) \\ \psi_{uk}(y, z) \end{pmatrix} := \begin{pmatrix} Q_k(1 - G(y|z)|z) \\ Q_k(G(y|z)|z) \end{pmatrix}$$

where $\tau, z \mapsto Q_k(\tau, z)$ is any conditional quantile function for the random variable x_k conditional on z and $y, z \mapsto G(y|z)$ is any conditional distribution function for the random variable y conditional on z . Denote by \mathcal{Q} and \mathcal{G} the parameter spaces for $\tau, z \mapsto Q_k(\tau|z)$, for all $k = 1, \dots, d_x$, and $y, z \mapsto G(y|z)$, respectively. Equip these spaces with norms $\|\cdot\|_{\mathcal{Q}}$ and $\|\cdot\|_{\mathcal{G}}$. Let further assume that C.1-C.4 hold.

Then, $\hat{\eta} - \eta_o = o_{P_o}(1)$ and $n_A^{1/2}(\hat{\eta} - \eta_o)$ converges in distribution to a random vector \mathbb{G}_o with a multivariate normal distribution.

Proof. To verify the conditions for the Law of Large Numbers in Lemma C.1 the Central Limit Theorem in Lemma C.2, we interpret the data $\{X_l\}_{l=1}^n$ as the two independent samples $\{y_i, z_i\}_{i=1}^{n_A}$ and $\{z_j, x_j\}_{j=n_A+1}^n$ described in Assumption D, and we assume that there is some number $0 < \kappa < 1$ such that, for $n_B = n - n_A$, as $n_A, n_B \rightarrow \infty$, $n_A/n \rightarrow \kappa$ and $n_B/n \rightarrow 1 - \kappa$. We set the true unknown finite parameter γ_o to

$$\gamma_o = (\mu_y^o, \mu_x^o, \mu_z^o, s_{yy}^o, s_{xx'}^o, s_{zz'}^o, s_{yz}^o, s_{xz}^o, e_{q,1}^o \lambda_{F1l}^o, \dots, e_{q,d_x}^o \lambda_{Fd_x u}^o)$$

where $\mu_y = E(y_i)$ and similarly for μ_x and μ_z . We set the finite dimensional parameter space \mathcal{A} equal to some compact subset \mathbb{K} of the Euclidean space \mathbb{R}^p , and we equip this space with the Euclidean norm. We equip Ψ with the sup-norm. In our case, the function $X_l, \gamma, \psi \mapsto m(X_l, \gamma, \psi) : \mathbb{R}^{1+d_x+d_z} \times \mathbb{R}^p \times \Psi \mapsto \mathbb{R}^p$ correspond to $(m_A(y_i, z_i, \gamma, \psi), m_B(z_j, x_j, \gamma), m_{AB}(z_l, \gamma), m_C(y_i, z_i, \gamma, \psi))'$ with:

$$m_A(y_i, z_i, \gamma) := \begin{pmatrix} \mu_y - y_i \\ s_{yy} - y_i^2 \\ s_{zy} - z_i y_i \end{pmatrix}$$

$$m_B(z_j, x_j, \gamma) := \begin{pmatrix} \mu_x - x_j \\ \text{vec}(s_{xx'}) - \text{vec}(x_j x_j') \\ \text{vec}(s_{xz'}) - \text{vec}(x_j z_j') \end{pmatrix}$$

$$m_{AB}(z_l, \gamma) := \begin{pmatrix} \mu_z - z_l \\ \text{vec}(s_{zz'}) - \text{vec}(z_l z_l') \end{pmatrix}$$

$$m_C(y_i, z_i, \gamma, \psi) := \begin{pmatrix} e_{q,1} \lambda_{F1l} - (q_{l\alpha} b + q_{l\beta} d) y_i Q_1(1 - G(y|z)|z) \\ e_{q_u,1} \lambda_{F1l} - (q_{u\alpha} b + q_{u\beta} d) y_i Q_1(G(y|z)|z) \\ \vdots \\ e_{q_{l,d_x}} \lambda_{Fd_x u} - (q_{l\alpha} b + q_{l\beta} d) y_i Q_{d_x}(1 - G(y|z)|z) \\ e_{q_{u,d_x}} \lambda_{Fd_x u} - (q_{u\alpha} b + q_{u\beta} d) y_i Q_{d_x}(G(y|z)|z) \end{pmatrix}$$

We set $M_n(\gamma, \psi)$ equal to

$$M_n(\gamma, \psi) := \begin{pmatrix} n_A^{-1} \sum_{i=1}^{n_A} m_A(y_i, z_i, \gamma) \\ n_B^{-1} \sum_{j=n_A+1}^n m_B(z_j, x_j, \gamma) \\ n^{-1} \sum_{l=1}^n m_{AB}(z_l, \gamma) \\ n_A^{-1} \sum_{i=1}^{n_A} m_C(y_i, z_i, \gamma, \psi) \end{pmatrix}$$

The estimator $\hat{\gamma} := \arg \min_{\gamma} \|M_n(\gamma, \hat{\psi})\|$ in our case corresponds to:

$$\hat{\gamma} = (\hat{\mu}_y, \hat{\mu}_x, \hat{\mu}_z, \hat{s}_{yy}, \hat{s}_{xx'}, \hat{s}_{zz'}, \hat{s}_{yz}, \hat{s}_{xz}, \hat{e}_{q,1} \hat{\lambda}_{F1l}, \dots, \hat{e}_{\bar{q},d_x} \hat{\lambda}_{Fd_x u})$$

To verify CLK Condition (1.1), it suffices to note that in our case $\|M_n(\hat{\gamma}, \hat{\psi})\|$ and $\inf_{\gamma} \|M_n(\gamma, \hat{\psi})\|$ are both equal to zero because there are no over-identifying restrictions (i.e., $p = k$). To verify CLK Condition (1.2), notice that $M(\gamma, \psi_o)$ delivers point identification of the finite dimensional parameter γ_o by assumption. To verify CLK Condition (1.3), notice that $\psi \mapsto M(\gamma, \psi)$ in our case is linear for all γ , and then continuous at $\psi = \psi_o$. We now verify CLK Condition (1.4). Because we have assumed that \hat{G} converges in probability to $G_{y|z}^o$ uniformly over y, z (see C.1) and that \hat{Q}_k converges in probability to $Q_{k|z}^o$ uniformly over τ, z (see C.2), it follows that $\hat{\psi}_k(y, z) = \left(\hat{Q}_k(1 - \hat{G}(y|z)|z), \hat{Q}_k(\hat{G}(y|z)|z) \right)$ converges in probability to $\psi_{ko}(y, z) = \left(Q_{k|z}^o(1 - G_{y|z}^o(y|z)|z), Q_{k|z}^o(G_{y|z}^o(y|z)|z) \right)$ uniformly over (y, z) . To verify CLK Condition (1.5'), we note that this condition is implied by CLK Condition (2.5'), which is verified below.

Hence, by the Law of Large Numbers in Lemma C.1., we have that $\hat{\gamma} - \gamma_o = o_P(1)$. Moreover, since η_o is a continuous function of γ_o , we can apply the Continuous Mapping Theorem (see e.g., van der Vaart, 1998, Theorem 18.11) to conclude that $\hat{\eta} - \eta_o = o_P(1)$.

We now verify CLK Conditions (2.1) - (2.6'). To verify that CLK Condition (2.1) is satisfied, we notice that in our case $\|M_n(\hat{\gamma}, \hat{\psi})\| = \inf_{\gamma} \|M_n(\gamma, \hat{\psi})\| = 0$ because there are no over-identifying restrictions. Verifying CLK Condition (2.2) is standard, so we omit it here. To verify CLK Condition (2.3), we start by noticing that, since $\psi \mapsto M(\gamma, \psi)$ is linear, the pathwise derivative exists and we have $\|M(\gamma, \psi) - M(\gamma, \psi_o) - \Gamma_2(\gamma, \psi_o)[\psi - \psi_o]\| = 0$. Hence, CLK Condition (2.3)(i) is satisfied. In our case, the difference $\Gamma_2(\gamma, \psi_o)[\psi - \psi_o] - \Gamma_2(\gamma_o, \psi_o)[\psi - \psi_o]$ is linear in $\gamma - \gamma_o$. This implies that any γ in the shrinking set \mathcal{A}_{δ_n} the sequence $\|\Gamma_2(\gamma, \psi_o)[\psi - \psi_o] - \Gamma_2(\gamma_o, \psi_o)[\psi - \psi_o]\|$ converges to zero as required by CLK Condition (2.3)(ii).

To verify CLK Condition (2.4), we start by noticing that $n^{1/4} \|\hat{\psi} - \psi_o\|_{\Psi}$ corresponds in our case to:

$$n_A^{1/4} \|\hat{\psi} - \psi_o\|_{\Psi} = n_A^{1/4} \sum_{k=1}^{d_x} \left\| \left(\hat{Q}_k(1 - \hat{G}), \hat{Q}_k(\hat{G}) \right) - \left(Q_{k|z}^o(1 - G_{y|z}^o), Q_{k|z}^o(G_{y|z}^o) \right) \right\|_{\Psi_k}$$

It then suffices to verify the condition for any given k . Add-and-subtract to the left hand side

$(Q_{k|z}^o(1 - \hat{G}), Q_{k|z}^o(\hat{G}))$ and apply the Triangle Inequality

$$\begin{aligned} n_A^{1/4} \|\hat{\psi}_k - \psi_{ko}\|_{\Psi} &\leq n_A^{1/4} \left\| \left(\hat{Q}_k(1 - \hat{G}), \hat{Q}_k(\hat{G}) - (Q_{k|z}^o(1 - \hat{G}), Q_{k|z}^o(\hat{G})) \right) \right\|_{\Psi} \\ &\quad + n_A^{1/4} \left\| \left(Q_{k|z}^o(1 - \hat{G}), Q_{k|z}^o(\hat{G}) - (Q_{k|z}^o(1 - G_{y|z}^o), Q_{k|z}^o(G_{y|z}^o)) \right) \right\|_{\Psi_k} \end{aligned}$$

C.2 implies that the first term in the right-hand-side of the latter display is $o_P(1)$. For some $\tilde{G} \in \mathcal{G}_{\delta_n}$, use the Inverse Function Theorem to obtain the following mean value expansion the second term in the right hand side:

$$n_A^{1/4} \|\hat{\psi}_k - \psi_{ko}\|_{\Psi} \leq o_P(1) + n_A^{1/4} \left\| \left(\frac{[G_{y|z}^o - \hat{G}]}{g_{k|z}^o(Q_{x|z}^o(1 - \tilde{G}))}, \frac{[\hat{G} - G_{y|z}^o]}{g_{k|z}^o(Q_{k|z}^o(\tilde{G}))} \right) \right\|_{\Psi_k}$$

Applying the Cauchy-Schwarz Inequality to the second term in the right-hand-side :

$$\begin{aligned} n_A^{1/4} \|\hat{\psi}_k - \psi_{ko}\|_{\Psi} &\leq o_P(1) \\ &\quad + \left\| \left((g_{k|z}^o(Q_{k|z}^o(1 - \tilde{G})))^{-1}, (g_{k|z}^o(Q_{k|z}^o(\tilde{G})))^{-1} \right) \right\|_{\Psi_k} \\ &\quad \times n_A^{1/4} \left\| (G_{y|z}^o - \hat{G}), (\hat{G} - G_{y|z}^o) \right\|_{\mathcal{G} \times \mathcal{G}} \end{aligned}$$

Under the assumption that the density $x \mapsto g_{k|z}^o(x|z)$ is bounded away from zero and bounded for all z (see Assumption C.3), we have that $\left\| \left((g_{k|z}^o(Q_{k|z}^o(1 - \tilde{G})))^{-1}, (g_{k|z}^o(Q_{k|z}^o(\tilde{G})))^{-1} \right) \right\|_{\Psi}$ is bounded and

$$n_A^{1/4} \|\hat{\psi}_k - \psi_{ko}\|_{\Psi} \leq o_P(1) + O(1)n_A^{1/4} \left\| (G_{y|z}^o - \hat{G}), (\hat{G} - G_{y|z}^o) \right\|_{\mathcal{G} \times \mathcal{G}}$$

The symmetry property of a norm and the convergence in probability in Assumption C.1 on the estimator of the distribution function implies that $n_A^{1/4} \left\| (G_{y|z}^o - \hat{G}), (\hat{G} - G_{y|z}^o) \right\|_{\mathcal{G} \times \mathcal{G}} = n_A^{1/4} \left\| (\hat{G} - G_{y|z}^o) \right\|_{\mathcal{G}} = o_P(1)$. Hence,

$$n_A^{1/4} \|\hat{\psi}_k - \psi_{ko}\|_{\Psi} \leq o_P(1) + O(1)o_P(1) \leq o_P(1)$$

To verify CLK Condition 2.5', we notice that the functions $\gamma \mapsto m_A(y_i, z_i, \gamma)$, $\gamma \mapsto m_B(z_j, x_j, \gamma)$ and $\gamma \mapsto m_{AB}(z_l, \gamma)$ are Holder continuous, and the function $\gamma, \psi \mapsto m_C(y_i, z_i, \gamma, \psi)$ is uniformly Lipchitz and twice continuously differentiable (because it is linear). Hence, a result in Chen, Linton and van Keilegom (2003, Theorem 3), implies that CLK Condition 2.5' (and 1.5') are satisfied.

Before verifying CLK Condition (2.6), we show that, under Assumption C.4, the difference $\hat{\psi}_k(y, z) - \psi_{ko}(y, z)$ has an asymptotic linear representation of the type:

$$\hat{\psi}_k(y, z) - \psi_{ko}(y, z) = \left(n_B^{-1} \sum_{j=n_A+1}^n \varphi_{Ql}(x_j, z_j, y, z) + \varphi_{Gl}(y, z) \right) + \left(R_{ln}(y, z) \right) \\ + \left(n_B^{-1} \sum_{j=n_A+1}^n \varphi_{Qu}(x_j, z_j, y, z) + \varphi_{Gu}(y, z) \right) + \left(R_{un}(y, z) \right)$$

with

$$\begin{aligned}
\varphi_{Ql}(x_j, z_j, y, z) &:= [\mathbf{1}(x_j \leq Q_{k|z}^o(1 - \hat{G}(y|z)|z))\varphi_Q(z_j) - 1 + \hat{G}(y|z)]/g_{x|z}^o(Q_{k|z}^o(1 - \hat{G}(y|z)|z)) \\
\varphi_{Qu}(x_j, z_j, y, z) &:= [\mathbf{1}(x_j \leq Q_{k|z}^o(\hat{G}(y|z)|z))\varphi_Q(z_j) - \hat{G}(y|z)]/g_{x|z}^o(Q_{k|z}^o(\hat{G}(y|z)|z)) \\
\varphi_{Gl}(y, z) &:= \left[g_{x|z}^o \left(Q_{x|z}^o \left(-n_A^{-1} \sum_{l=1}^{n_A} [\mathbf{1}(y_l \leq y)\varphi_G(z_l) - G_{y|z}^o(y|z)] \right) \middle| z \right) \right]^{-1} \\
\varphi_{Gu}(y, z) &:= \left[g_{x|z}^o \left(Q_{x|z}^o \left(n_A^{-1} \sum_{l=1}^{n_A} [\mathbf{1}(y_l \leq y)\varphi_G(z_l) - G_{y|z}^o(y|z)] \right) \middle| z \right) \right]^{-1},
\end{aligned}$$

and, for $s = l, u$, $R_{sn}(y, z) = o_{P_o}(n_A^\delta)$ for some $\delta \leq 1/2$ uniformly over (y, z) . For $\hat{\psi}_{ku}(y, z) - \psi_{kuo}(y, z) := \hat{Q}_k(\hat{G}(y|z)|z) - Q_{k|z}^o(G_{y|z}^o(y|z)|z)$, add-and-subtract $Q_{k|z}^o(\hat{G}(y|z)|z)$ to obtain:

$$\hat{\psi}_{ku}(y, z) - \psi_{kuo}(y, z) = \hat{Q}_k(\hat{G}(y|z)|z) - Q_{k|z}^o(\hat{G}(y|z)|z) + Q_{k|z}^o(\hat{G}(y|z)|z) - Q_{k|z}^o(G_{y|z}^o(y|z)|z)$$

Under Assumption (C.4), the asymptotic linear expansion for $\hat{G} - G_{y|z}^o$ and the Functional Delta Method (see e.g., Van der Vaart, 1998, Theorem 20.8) implies that the last difference in the right-hand-side term can be replaced by:

$$\hat{\psi}_{ku}(y, z) - \psi_{kuo}(y, z) = \hat{Q}_k(\hat{G}(y|z)|z) - Q_{k|z}^o(\hat{G}(y|z)|z) + \varphi_{Gu}(y, z) + o_{P_o}(n_A^\delta)$$

Furthermore, the asymptotic linear expansion for $\hat{Q}_k - Q_{k|z}^o$ in Assumption A.B.1.4 implies that the difference in the right-hand-side can be replaced by:

$$\begin{aligned}
\hat{\psi}_{ku}(y, z) - \psi_{kuo}(y, z) &= n_B^{-1} \sum_{j=n_A+1}^n \varphi_{Qu}(x_j, z_j, y, z) + o_{P_o}(n_B^\delta) + \varphi_{Gu}(y, z) + o_{P_o}(n_A^\delta) \\
&= n_B^{-1} \sum_{j=n_A+1}^n \varphi_{Qu}(x_j, z_j, y, z) + \varphi_{Gu}(y, z) + R_{ln}(y, z)
\end{aligned}$$

$R_{ln}(y, z) = o_{P_o}(n_A^\delta) = o_{P_o}(n_B^\delta)$ uniformly over y, z . Obtaining the asymptotic linear expansion for $\hat{\psi}_{kl}(y, z) - \psi_{klo}(y, z) := \hat{Q}_k(1 - \hat{G}(y|z)|z) - Q_{k|z}^o(1 - G_{y|z}^o(y|z)|z)$ is similar and thus omitted. To verify CLK Condition (2.6), we start from:

$$\Gamma_2(\gamma_o, \psi_o)[\psi - \psi_o] = \begin{pmatrix} -e_{\underline{q}}^o E \left(y [Q_k(1 - G(y|z)|z) - Q_{k|z}^o(1 - G_{y|z}^o(y|z)|z)] \right) \\ \vdots \\ -e_{\underline{q}}^o E \left(y [Q_k(G(y|z)|z) - Q_{k|z}^o(G_{y|z}^o(y|z)|z)] \right) \end{pmatrix}$$

We now substitute in the uniform asymptotic linear representation for $\hat{\psi}(y, z) - \psi_o(y, z)$ under C.4, interchange integral and summation, apply the law of total expectations, and approximate the unconditional expectation to obtain

$$\Gamma_2(\gamma_o, \psi_o)[\hat{\psi} - \psi_o] = \begin{pmatrix} n_A^{-1} \sum_{i=1}^{n_A} -e_{\underline{q}}^o \varphi_l^*(y_i, z_i) \\ \vdots \\ n_A^{-1} \sum_{i=1}^{n_A} -e_{\underline{q}}^o \varphi_u^*(y_i, z_i) \end{pmatrix} + o_{P_o}(n_A^{-1/2}) + \begin{pmatrix} n_A^{-1} \sum_{i=1}^{n_A} -e_{\underline{q}}^o R_{nl}(y_i, z_i) \\ \vdots \\ n_A^{-1} \sum_{i=1}^{n_A} -e_{\underline{q}}^o R_{nu}(y_i, z_i) \end{pmatrix}$$

with

$$\begin{aligned}\varphi_l^*(y_i, z_i) &= E\left[y_i\left(n_B^{-1} \sum_{j=n_A+1}^n \varphi_l(x_j, z_j, y_i, z_i) + \varphi_{Gl}(y_i, z_i)\right) \middle| y_i, z_i\right] \\ \varphi_u^*(y_i, z_i) &= E\left[y_i\left(n_B^{-1} \sum_{j=n_A+1}^n \varphi_u(x_j, z_j, y_i, z_i) + \varphi_{Gu}(y_i, z_i)\right) \middle| y_i, z_i\right]\end{aligned}$$

Averaging the remainder terms and using the definition of $M_n(\gamma_o, \psi_o)$, it follows that

$$M_n(\gamma_o, \psi_o) + \Gamma_2(\gamma_o, \psi_o)[\hat{\psi} - \psi_o] = o_P(n^{-1/2}) + o_P(n^{-1/2}) = o_P(n^{-1/2})$$

which is, under standard conditions for sum of independent random variables with finite variance (see e.g., Van der Vaart, 1998, Proposition 2.27), asymptotically normal with mean zero and finite variance, as required in CLK Condition (2.6).

Hence, by the Central Limit Theorem in Lemma C.2., we have that $(\hat{\gamma} - \gamma_o)n_A^{1/2}$ converges in distribution to a random vector with multivariate normal distribution. Since $n_A^{1/2}(\hat{\eta} - \eta_o)$ is a differentiable function of $n_A^{1/2}(\hat{\gamma} - \gamma_o)$, we conclude from the Delta Method that $n_A^{1/2}(\hat{\eta} - \eta_o)$ converges in distribution to some random vector \mathbb{G}_o with multivariate normal distribution. ■

We next provide conditions under which the non-parametric bootstrap can consistently estimate the asymptotic distribution of $(\hat{\eta} - \eta_o)n_A^{1/2}$. To obtain such conditions, we rely again on a Theorem by Chen, Linton and Van Keilegom (2003, Theorem B), which is re-stated below for convenience:

Lemma C.4 (*Consistency of the Non-Parametric Bootstrap - Chen, Linton, Van Keilegom, 2003, Theorem B*). *Let $\hat{\psi}^*$ be the same estimator as $\hat{\psi}$ but based on bootstrap data. Here, and subsequently, superscript \star denotes a moment computed under the bootstrap distribution conditional on the original data. Define the bootstrap estimator $\hat{\gamma}^* := \|M_n^*(\hat{\theta}^*, \hat{\psi}^*) - M_n(\hat{\gamma}, \hat{\psi})\|$. Suppose that $\{X_l\}_{l=1}^n$ is i.i.d.; $\gamma \in \text{int}(\mathcal{A})$; that CLK Conditions (2.1), (2.4), (2.5') and (2.6) hold with 'in probability' replaced by 'almost sure'; that CLK Conditions (2.2) holds with ψ_o replaced by $\psi \in \Psi_{\delta_n}$ while CLK Condition (2.3) holds with ψ_o replaced by $\psi \in \Psi_{\delta_n}$; and that $\gamma, \psi \mapsto \Gamma_1(\gamma, \psi)$ is continuous in ψ at $\gamma = \gamma_o, \psi = \psi_o$. Suppose:*

(2.4B) *With P^* -probability tending to one, $\hat{\psi}^* \in \Psi$, and $\|\hat{\psi}^* - \hat{\psi}\|_{\Psi} = o_{P^*}(n^{-1/4})$.*

(2.5'B) $\sup_{(\gamma, \psi) \in \mathcal{A}_{\delta_n} \times \Psi_{\delta_n}} \|M_n^*(\gamma, \psi) - M_n(\gamma, \psi) + M_n^*(\gamma_o, \psi_o) - M_n(\gamma_o, \psi_o)\| = o_{P^*}(n^{-1/2})$ for all positive values $\delta_n = o(1)$.

(2.6B) $\sqrt{n}\|M_n^*(\hat{\gamma}, \hat{\psi}) - M_n(\hat{\gamma}, \hat{\psi}) + \Gamma_2(\hat{\gamma}, \hat{\psi})[\hat{\psi}^* - \hat{\psi}]\|$ converges in distribution to a random vector with multivariate normal distribution.

Then, $(\hat{\gamma} - \gamma_o)\sqrt{n}$ converges in probability to $(\hat{\gamma}^* - \hat{\gamma})\sqrt{n}$ in P^* -probability.

CLK Conditions (2.4B)-(2.6'B) can be verified under the same assumptions implying CLK Conditions (2.4) and (2.6'), by using the corresponding asymptotic linear approximation for $\hat{\psi}^* - \hat{\psi}$. This observation establishes the consistency of the non-parametric bootstrap to approximate the distribution of the random vector \mathbb{G}_o . Following the results in Chen, Linton and Van Keilegom (2003, Theorem 2), we could derive a closed-form expression for the variance of \mathbb{G}_o . Such an expression however is unnecessary to prove the validity of the inference procedures in the text.

APPENDIX D: SENSITIVITY TO CHOICE OF TUNING PARAMETERS

In this section, we investigate the sensitivity to choice of tuning parameters on the conclusions derived from the Monte Carlo in Section 5.

Implementing the series non-parametric estimator of the conditional survival, distribution, and quantile functions in the Frechet, Analog, Bias-C and Before estimators requires to choose the location and number of knots. Following standard practice in series estimation, we place the knots at the quantiles of z . Choosing the number of knots according to the function $K_n = L_n = \lfloor n_B^{1/3} \rfloor$ is less clear. To evaluate the sensitivity of the results to this choice, we make vary the numbers of knots K_n and L_n across experiments according to the function $K_n = L_n = \lfloor n_B^r \rfloor$ for $r \in \{1/9, 1/7, 1/5, 1/3\}$. Roughly speaking, estimators with number of knots below $\lfloor n_B^{1/5} \rfloor$ will tend to oversmooth the underlying target function while estimator with number of knots above $\lfloor n_B^{1/5} \rfloor$ will tend to undersmooth it. Table 5 below reports the root mean squared error uniformly integrated (RMSEI) for different choices of knots in the normal and non-normal experiments. The results suggest that the choice of the number of knots has a non-negligible effect on the performance of the estimator. Oversmoothing may result in estimates with larger RMSEI than estimates obtained with undersmoothing. Since the choice of the number of knots has an important effect on the performance of the estimator, it would be useful to have a way to choose them in practice. Obtaining such a result is however beyond the scope of this paper.

The number of Monte Carlo replications in the experiments reported in Section 5 is 250 and the number of bootstrap replications is 250. To evaluate the sensitivity of result to this choice, we have run 1,000 replications with 500 bootstrap replications for the experiments with the smaller sample size (i.e., $n = 250$). Results are reported in Tables 6 and 7. The qualitative conclusions obtained from 250 Monte Carlo and bootstrap replications remain unchanged.

TABLE 5. *Sensitivity of Estimators to the First Step Non-Parametric Estimator.*
Normal and Non-Normal Experiments

Experiment	Obs.	Estimator	Knots	Covariate z ($\beta_2 = 1$)				Covariate x ($\alpha_1 = 1$)					
				Dec.	AL	ASE	RMSEI	Dec.	AL	ASE	RMSEI		
Normal	250	Frechet	2	.042	.025	.018	.294	1.00	.598	.006	1.26		
			3	.042	.025	.018	.293	1.00	.587	.006	1.26		
			5	.041	.024	.018	.290	.988	.563	.005	1.24		
		Analog	2	.042	.025	.018	.293	1.00	.592	.005	1.26		
			3	.042	.024	.018	.292	.998	.582	.004	1.25		
			5	.041	.024	.018	.290	.990	.562	.004	1.24		
		Bias-C(.5)	2	.042	.019	.018	.285	1.00	.573	.005	1.25		
			3	.042	.019	.018	.284	.998	.564	.005	1.25		
			5	.041	.019	.018	.281	.988	.544	.005	1.24		
		Bias-C(1)	2	.043	.015	.018	.277	1.00	.555	.005	1.25		
			3	.042	.014	.018	.276	.997	.545	.005	1.24		
			5	.042	.014	.018	.274	.987	.526	.005	1.23		
		Non- Normal	250	Frechet	2	.004	.007	.133	.381	.876	.783	.200	1.36
					3	.004	.007	.130	.377	.890	.773	.167	1.35
					5	.004	.006	.129	.375	.850	.750	.184	1.33
Analog	2			.005	.005	.122	.365	.872	.594	.052	1.23		
	3			.005	.005	.121	.364	.878	.600	.048	1.23		
	5			.005	.005	.120	.362	.880	.598	.053	1.23		
Bias-C(.5)	2			.005	.004	.138	.384	.893	.601	.055	1.24		
	3			.005	.004	.137	.382	.901	.607	.051	1.24		
	5			.005	.004	.135	.381	.900	.605	.054	1.24		
Bias-C(1)	2			.005	.003	.156	.405	.914	.608	.061	1.25		
	3			.005	.003	.154	.403	.924	.614	.057	1.25		
	5			.005	.003	.152	.401	.920	.612	.059	1.26		

This table presents different measures describing the finite sample performance of confidence intervals for the coefficients of in the Normal Experiment. The label "Obs." indicates the number of observations in each sample. " ρ_{zx} " stands for the correlation between x and z . The number of bootstrap replications is 250. The number of Monte Carlo replications is 250.

TABLE 6. *Sensitivity of Estimators to the Number of Monte Carlo and Bootstrap Replications.*
Normal and Non-Normal Experiments

Experiment	Obs.	Estimator	Covariate z ($\beta_2 = 1$)				Covariate x ($\alpha_1 = 1$)			
			Dec.	AL	ASE	RMSEI	Dec.	AL	ASE	RMSEI
Normal	250	Matrix	.043	.028	.017	.298	1.00	.671	.002	1.29
		Frechet	.043	.023	.016	.289	.992	.566	.005	1.25
		Analog	.043	.023	.016	.289	.992	.564	.004	1.24
		Bias-C(.5)	.044	.018	.016	.282	.992	.545	.005	1.24
		Bias-C(1)	.044	.014	.017	.275	.992	.527	.005	1.23
		Before	.045	.012	.016	.273	1.09	.522	.004	1.27
Non- Normal	250	Matrix	.007	.006	.106	.347	.947	.677	.007	1.27
		Frechet	.006	.006	.120	.365	.858	.746	.310	1.38
		Analog	.007	.005	.111	.351	.885	.592	.103	1.25
		Bias-C(.5)	.006	.004	.134	.381	.884	.603	.227	1.30
		Bias-C(1)	.006	.003	.166	.420	.883	.614	.520	1.42

This table presents different measures describing the finite sample performance of confidence intervals for the coefficients of in the Normal Experiment. The label "Obs." indicates the number of observations in each sample. " ρ_{zx} " stands for the correlation between x and z . The number of bootstrap replications is 500. The number of Monte Carlo replications is 1000.