Identifying Structural Effects in Nonseparable Systems Using Covariates

Halbert White[*] Karim Chalak
UC San Diego Boston College

October 16, 2008

Abstract This paper demonstrates the extensive scope of an alternative to standard instrumental variables methods, namely covariate-based methods, for identifying and estimating effects of interest in general structural systems. As we show, commonly used econometric methods, specifically parametric, semi-parametric, and nonparametric extremum or moment-based methods, can all exploit covariates to estimate well-identified structural effects. The systems we consider are general, in that they need not impose linearity, separability, or monotonicity restrictions on the structural relations. We consider effects of multiple causes; these may be binary, categorical, or continuous. For continuous causes, we examine both marginal and non-marginal effects. We analyze effects on aspects of the response distribution generally, defined by explicit or implicit moments or as optimizers (e.g., quantiles). Key for identification is a specific conditional exogeneity relation. We examine what happens in its absence and find that identification generally fails. Nevertheless, local and near identification results hold in its absence, as we show.

JEL Classification Numbers: C10, C20, C30, C50.

Keywords: conditional exogeneity; extremum estimator; identification; method of moments; nonparametric estimation; structural equations.

[*] Corresponding author. Address: Department of Economics 0508, UCSD, La Jolla, CA 92093-0508. Phone: 858 534-3502; Fax 858 534-7040; Email: hwhite@ucsd.edu

1 Introduction

Classical systems of structural equations are parametric, linear, and separable. Beginning in the 1980's considerable attention has been directed toward relaxing some or all of these conditions. For example, Brown (1983) studies parametric separable systems nonlinear only in regressors. Newey and Powell (1988, 2003), Newey, Powell, and Vella (1999), Darolles, Florens, and Renault (2003), Das (2005), Hall and Horowitz (2005), and Santos (2006), among others, study nonparametric separable systems. Roehrig (1988), Brown and Matzkin (1998), Chesher (2003, 2005), Imbens and Newey (2003), Matzkin (2003, 2004, 2005), Chernozhukov and Hansen (2005), and Chernozhukov, Imbens, and Newey (2007), for example, consider nonparametric nonseparable systems but impose monotonicity conditions at some stage in their analysis. Angrist and Imbens (1994), Angrist, Imbens, and Rubin (1996), Heckman (1997), Heckman, Ichimura, and Todd (1998), Heckman and Vytlacil (1999, 2001, 2005, 2007), Blundell and Powell (2003), Heckman, Urzua, and Vytlacil (2006), and Hahn and Ridder (2007), among others, study the consequences of otherwise relaxing the classical assumptions, often considering systems with binary treatments or index structures.

In addition, researchers have increasingly focused attention on aspects of the distribution of the response of interest beyond the conditional mean. An area of particular interest is on structural modeling of response quantiles. Examples are the work of Chesher (2003, 2005), Imbens and Newey (2003), and Chernozhukov and Hansen (2005).

Recently, Altonji and Matzkin (2005), Hoderlein (2005, 2007), Hoderlein and Mammen (2007), and Schennach, White, and Chalak (2007) have analyzed structural systems in which the structural equations may be nonparametric, nonlinear, and nonseparable, without necessarily imposing monotonicity. Because economic theory is often not fully specific about the forms of the relationships of interest, this generality should enhance economists' ability to study economic relationships without having to make assumptions that might not be supported by the data. Moreover, results for the general case provide a foundation for testing restrictions suggested by economic theory.

As is evident from the literature cited, and as Darolles, Florens, and Renault (2003) and Chalak and White (2007a) (CW) discuss, there are a variety of ways to identify structurally meaningful features of interest in non-classical contexts. The classical method is that of instrumental variables (IV), which requires the availability of exogenous instruments, that is, variables uncorrelated with or independent of the unobserved causes of the response of interest. Schennach, White, and Chalak (2007) analyze the properties of effect estimators based on such instruments in the general nonseparable context.

An alternative to the use of exogenous instruments to identify and estimate structural effects involves the use of covariates. This method originates from the treatment effects literature (e.g., Rubin, 1974; Rosenbaum and Rubin, 1983) and has been significantly developed and enhanced by Barnow, Cain, and Goldberger (1980), Heckman and Robb (1985), Heckman, Ichimura, and Todd (1998), Hahn (1998), Hirano and Imbens (2001, 2004), Hirano, Imbens, and Ridder (2003), and Heckman and Vytlacil (2005), among others. In contrast to standard exogenous instrumental variables, covariates are typically *endogenous*. Covariates operate by ensuring the *conditional* exogeneity of the causes of interest. That is, conditional on the covariates, the causes of interest are independent of the relevant unobservables, making possible identification and estimation of effects of interest. Because of their role as conditioning variables and because they are instrumental in identification and estimation, we can view covariates as *conditioning instruments* (see CW).

In particular, Altonji and Matzkin (2005), Hoderlein (2005, 2007), and Hoderlein and Mammen (2007) use covariates to identify structural features of interest in nonseparable systems. They have also been used in quantile regression (e.g., Firpo, 2007). Imbens and Newey (2003) discuss their use in a variety of contexts. Although this suggests that the scope of covariate-based methods for identifying and estimating structural effects is considerable, the full extent of this scope is as yet an open question.

A main contribution of this paper is thus to demonstrate the extensive scope of covariate-based methods for identifying and estimating effects of interest in general structural systems. As we show, all of the procedures commonly used in econometrics, for example parametric, semi-parametric, and nonparametric extremum or moment-based methods, can exploit covariates to estimate well-identified structural effects. Moreover, these results hold for general structural systems, without imposing linearity, separability, or monotonicity.

We analyze the use of covariates to identify structural effects of interest, generally defined. In this context, identification means equality between the structural effect of interest and a well defined stochastic object (e.g., one involving a conditional mean or quantile) that can be defined solely in terms of observables. Once this form of identification has been established, then appropriate statistical methods can be deployed to estimate the stochastic object, and thus the effect of interest.

Specifically, we examine the simultaneous identification of structural effects of multiple causes; these may be binary, categorical, or continuous. For the case of continuous causes, we examine both marginal and non-marginal effects. In section 2, we consider average effects, taking care to rigorously define these effects and to distinguish structurally meaningful objects from stochastically meaningful objects. Section 3 analyzes more general effects, based on explicit moments, implicit moments, or aspects of the response distribution defined as optimizers (e.g., quantiles). To the best of our knowledge, the use of implicit moment conditions to define effects of interest is new.

A key condition ensuring the identification of effects of interest is a specific conditional exogeneity relation. In section 4, we study the role of this condition by examining what happens in its absence, finding that identification then generally fails. Neverthless, we provide new local and near identification results that provide insight into how the various effect measures are impacted by departures from conditional exogeneity and by the sensitivity of the response of interest to unobservables.

Section 5 concludes with a summary and discussion of directions for further research.

2 Data Generation and Average Effects

2.1 Data Generation

We first specify a recursive data generating process. In recursive systems, there is an inherent ordering of the variables: "predecessor" variables may determine "successor" variables, but not vice versa. For example, when X determines Y, then Y cannot determine X. In such cases, we say that Y *succeeds* X, and we write $Y \Leftarrow X$ as a shorthand notation. Throughout, random variables are defined on $(\Omega, \mathcal{F}, P)$, a complete probability space.

Assumption A.1 (i) Let $(D', U', V', W', Y, Z')'$ be a vector of random variables such that Y is scalar valued, D, W, and Z are finitely dimensioned, and U and V are countably dimensioned; (ii) Suppose further that a recursive structural system generates (D, W, Y, Z) such that $Y \Leftarrow (D, U, V, W, Z)$, $D \Leftarrow (U, V, W, Z)$, $W \Leftarrow (U, V, Z)$, and $Z \Leftarrow U, V$, with structural equations

$$Y = r(D, Z, U), \qquad D = q(Z, W, U, V),$$

where the response functions q and r are unknown measurable functions mapping to $\mathbb{R}^k$, $k \in \mathbb{N}$, and $\mathbb{R}$, respectively; and (iii) the realizations of D, W, Y, and Z are observed, whereas those of U and V are not.

Assumption A.1(i) imposes only stochastic structure. No structural relations need hold under A.1(i); these are imposed by A.1(ii). We emphasize that by specifying that D, W, Y, and Z are structurally generated, A.1(ii) embodies systems in which the determining equations represent economic structure arising, for example, from optimizing behavior and/or equilibrium. As Goldberger (1972, p.979) points out, such structural relations are *directional* causal links. For example, variations in D, Z, and U cause variations in Y, whereas variations in Y have no impact on D, Z, and U. The special nature of these relations has been clearly articulated by Strotz and Wold (1960) and Fisher (1966, 1970); see also CW.

Our interest attaches to the effects of one or more elements of D on Y. Accordingly, we call Y the *response of interest* and D *causes of interest.* We call the remaining observables $X := (W, Z)$ *covariates.* Nothing in A.1 ensures that D, W, or Z is independent of U, so D, W, and Z are generally endogenous.

Whereas Imbens and Newey (2003) and (implicitly) Hoderlein (2005, 2007) study a recursive structure in which q is monotonic (or even additively separable) in a scalar unobservable, no such requirements are imposed here. Specifically, we do not assume q or r to be linear, separable, or monotone in their arguments. Significantly, the unobservables U and V can be vectors; these vectors can even be of infinite dimension. This permits wide latitude in accommodating variables known to drive the response or causes of interest, but which cannot be observed. Aspects of this flexibility can also be found in work of

Altonji and Matzkin (2005), Hoderlein and Mammen (2007), and Shennach, White, and Chalak (2007).

Given that $D \Leftarrow (U, V, W, Z)$, there is no restriction in assuming $D = q(Z, W, U, V)$. We make this explicit for clarity and convenience; for succinctness, we do not make explicit the structural equations for W and Z. We consider some restrictions on q below, following Proposition 2.3. Although q may be a function of W, it need not be; generally, there will be elements of W that do not drive D. In contrast, we explicitly exclude W from r. This loses no generality, as W may have dimension zero. Nevertheless, when W has positive dimension, it can play an instrumental role in identifying effects of interest (cf. Altonji and Matzkin, 2005). Excluding V from r rules out V as a direct cause of Y and distinguishes V from U. This exclusion can assist in identification, as Proposition 2.3 below suggests. V may also be of dimension zero, however.

When structurally meaningful objects of interest can be equated to well-defined and empirically accessible standard stochastic objects, we say they are *identified*. The most straightforward such stochastic object relevant here is the conditional expectation $\mu(D, X) := E(Y \mid D, X)$. To begin to investigate the content of $\mu(D, X)$, we let $\text{supp}(D, X)$ denote the support of (D, X), the smallest measurable set on which the density of (D, X), say dH, integrates to one. Also, let $dG(u \mid d, x)$ define the conditional density of U given (D, X).

We begin with a simple result relating μ to the structure determining Y.

Proposition 2.1 *Suppose Assumption A.1(i) holds with* $E(Y) < \infty$*. Then* (i) $\mu(D, X) := E(Y \mid D, X)$ *exists and is finite; if in addition A.1(ii) holds, then* (ii) *for each* (d, x) *in* $supp(D, X)$

$$\mu(d, x) = \int r(d, z, u)\, dG(u \mid d, x). \qquad \blacksquare$$

Part (i) ensures that $\mu(D, X)$ is a well defined stochastic object without requiring any particular causal structure. When the structure of part (ii) holds, we can represent $\mu(d, x)$ as the *average response given* $(D, X) = (d, x)$. For this, we require the conditional density $dG(\,\cdot \mid d, x)$ to be regular, as defined in Dudley (2002, ch.10.2). We assume throughout that any referenced conditional densities are regular. We call μ an *average response function*. This tells us the expected response given realizations (d, x) of (D, X).

Nevertheless, μ does not necessarily provide insight into the effects of D on Y, as such effects require knowledge about what happens under interventions to D, that is, under variations in the realized value d unrelated to the stochastic behavior of X or any other random variable of the system.

We specify the expected response under interventions by defining the *average counterfactual response at d given $X = x$*. When $E(r(d, Z, U))$ exists and is finite for each d in supp(D), this is

$$\rho(d \mid x) := E(r(d, Z, U) \mid X = x) = \int r(d, z, u)\, dG(u \mid x),$$

where $dG(\cdot \mid x)$ is the conditional density of U given $X = x$. The term "counterfactual" is used here to signal that the value d is arbitrary and need not correspond to the stochastic behavior specified in A.1(i) or the structural behavior specified in A.1(ii).

The notation $\rho(d \mid x)$ is intended to emphasize the difference between the roles played by d and x. Whereas $X = x$ is *given* in the usual sense of stochastic conditioning, D is *set* to d by intervention. That is, the value of D is not determined by the structure of A.1, but is instead set to any value in supp (D), as in Strotz and Wold (1960). Pearl (1995, 2000) introduced the "do" operator to express such counterfactual settings. When D is set to d, Pearl writes the expected response given $X = x$ as $E(Y \mid \text{do}(d), X = x)$. In our notation, $E(Y \mid \text{do}(d), X = x) = \rho(d \mid x)$.

The function ρ is a conditional analog of Blundell and Powell's (2003) "average structural function." To make clear the counterfactual nature of ρ, we call it a *covariate-conditioned counterfactual average response function* or a "counterfactual average response function," leaving conditioning implicit.

The assumption that $D \Leftarrow (U, V, W, Z)$ ensures that when D is set to different values for d, this does not necessitate different realizations for (U, V, W, Z). We thus view $\rho(d \mid x)$ as representing $\rho(d \mid X(\omega))$ for $X(\omega) = x$, where X explicitly does not depend on d. Otherwise, one must consider

$$\rho(d \mid X(d, \cdot)) = E(r(d, Z(d, \cdot), U(d, \cdot) \mid X(d, \cdot)),$$

making the dependence of (U, W, Z) (hence X) on d explicit. This permits analysis of mediated effects, but because this is somewhat involved, we leave this aside here (see CW, sections 4.1.2 and 4.2.3). By assuming that $D \Leftarrow (U, V, W, Z)$, we ensure that $\rho(d \mid x)$ gives a representation in which d and x are variation-free.

Comparing $\mu(d, x)$ and $\rho(d \mid x)$, we see that $dG(u \mid d, x)$ appears in $\mu(d, x)$, whereas $dG(u \mid x)$ appears in $\rho(d \mid x)$. If for given x and all d, $dG(\,\cdot \mid d, x) = dG(\,\cdot \mid x)$, then $\mu(d, x) = \rho(d \mid x)$. If this holds for all (d, x) in supp (D, X), then we write $\mu = \rho$. When, as happens here, a stochastically meaningful object like μ is identified with a structurally meaningful object, we say that it is *structurally identified*. Similarly, when a structurally meaningful object like ρ is identified with a stochastic object, we say it is *stochastically identified*. If stochastic identification holds uniquely with a representation solely in terms of observable variables, we say that both the stochastic object and its structural counterpart are *fully identified*. These definitions conform to CW and Schennach, White, and Chalak (2007).

The condition that $dG(\,\cdot \mid d, x) = dG(\,\cdot \mid x)$ for all (d, x) in supp (D, X) is a conditional independence requirement, analogous to that imposed in similar contexts by Altonji and Matzkin (2005, assumption 2.1), Hoderlein (2005, 2007), and Hoderlein and Mammen (2007). Here we use Dawid's (1979) conditional independence notation $D \perp U \mid X$ to denote that D is independent of U given X.

Assumption A.2 $D \perp U \mid X$.

By analogy with the use of "exogeneity" to describe regressors independent of unobservable "disturbances" (e.g., Wooldridge, 2002, p. 50), when A.2 holds we say D is *conditionally exogenous* (cf. CW). This concept involves only the data generating process and does not involve any parametric model; it is thus distinct from weak, strong, or super exogeneity (Engle, Hendry, and Richard, 1983), which are defined in terms of the properties of parametric models. It contains strict exogeneity ($D \perp U$) as a special case (when $X \equiv 1$). Following CW, we call X "conditioning instruments," in recognition of the instrumental role X plays in identifying μ with ρ. When the covariates suffice to ensure conditional exogeneity for D, we call them *sufficient covariates*, following Dawid (1979). Given A.1, conditional exogeneity ensures the natural generalization of Rosenbaum and Rubin's (1983) "unconfoundedness" condition introduced by Hirano and Imbens (2004): for all d in supp (D), $Y_d \perp D \mid X$, where $Y_d := r(d, Z, U)$.

Whereas Altonji and Matzkin (2005) assume the covariates X do not enter r (so our Z is null), Hoderlein and Mammen (2007) assume that all covariates enter r (so our W is null). Assumption A.2 contains these possibilities as special cases; neither restriction is necessary here.

Our first identification result formalizes this discussion.

Theorem 2.2 *Suppose A.1(i,ii) and A.2 hold. (i.a) Then for all $(d, x) \in supp\ (D, X)$, $\rho(d \mid x) := \int r(d, z, u)\, dG(u \mid x)$ exists and $\rho(d \mid x) = \mu(d, x)$, that is, $\rho = \mu$, so that ρ is stochastically identified and μ is structurally identified. (i.b) If in addition A.1(iii) holds, then ρ and μ are fully identified.* ■

Note that this result ensures that ρ is defined on supp(D, X); outside this set, we leave ρ undefined. In subsequent results we encounter similar situations, and we will understand the functions in those results to be defined only on the specified support. When we refer to ρ or its analogs below, we will understand that attention is restricted to the support on which the function of interest is defined.

As we see in Section 4, when A.2 fails, so does full identification. A.2 is thus a crucial identifying assumption. The specific structure relating D, U, V, W, and Z determines whether Assumption A.2 holds or is plausible, as discussed in CW and Chalak and White (2007b,c). For example, A.2 holds for Pearl's (1995, 2000) "back door" structures and for structures in which W acts as a vector of predictive proxies for U, as in White (2006). The next result provides conditions ensuring A.2.

Proposition 2.3 *Suppose that A.1(i) holds. Then $U \perp (D, V) \mid X$ if and only if $U \perp V \mid X$ and $U \perp D \mid (V, X)$.* ■

Thus, $U \perp V \mid X$ and $U \perp D \mid (V, X)$ ensure that $D \perp U \mid X$. In particular, given A.1(ii), if U is not a direct cause of D, so that $D = q(Z, W, V)$, then $U \perp D \mid (V, X)$. Thus, if we also have $U \perp V \mid X$, then $D \perp U \mid X$. In this case U and V are distinct sources of unobserved variation for Y and D, respectively. Alternatively, suppose U directly causes D and that V is the sole direct cause of U. Then D is measurable-$\sigma(X, V)$, and we again have $U \perp D \mid (V, X)$. Again, it suffices for A.2 that $U \perp V \mid X$, although this can hold only for certain specific structural relations between U and V.

2.2 Average Effect Measures

Using ρ, we can define the expected effect of any intervention $d \to d^* := (d, d^*)$ to D. Specifically, the *average effect on* Y *of the intervention* $d \to d^*$ *to* D *given* $X = x$ is

$$\Delta\rho(d, d^* \mid x) := \rho(d^* \mid x) - \rho(d \mid x).$$

We call this a "covariate-conditioned average effect," or, leaving conditioning implicit, an "average effect." As a special case, this includes the conditional average treatment effect of a binary treatment discussed by Abadie and Imbens (2002). To ensure that this effect is well defined, we require that (d, d^*, x) is "admissible," that is, that $(d, d^*) \in \text{supp}(D) \times \text{supp}(D)$ and that $x \in \text{supp}(X \mid D = d) \cap \text{supp}(X \mid D = d^*)$, where $\text{supp}(X \mid D = d)$ is the support of X given that $D = d$. Formally, the *interventions* $(\omega, \omega^*) \in \Omega \times \Omega$ *underlying* $d \to d^*$ *given* $X = x$ are those (ω, ω^*) pairs satisfying $d = D(\omega)$, $x = X(\omega)$ and $d^* = D(\omega^*)$, $x = X(\omega^*)$.

When ρ is stochastically identified, then for all admissible (d, d^*, x), we have

$$\Delta\rho(d, d^* \mid x) = \Delta\mu(d, d^*, x) := \mu(d^*, x) - \mu(d, x).$$

When μ is fully identified, a consistent estimator of $\Delta\mu(d, d^*, x)$ provides a consistent estimator of $\Delta\rho(d, d^* \mid x)$.

Replacing x with X yields a random version of the average effect, $\Delta\rho(d, d^* \mid X)$, with an optimal prediction property. Specifically, by the mean-square optimality of conditional expectation, $\Delta\rho(d, d^* \mid X)$ is the mean squared error-best predictor of the random effect $\Delta r(d, d^*, Z, U) := r(d^*, Z, U) - r(d, Z, U)$ among all predictors based on X.

This refines the usual ceteris paribus interpretation of effects: $\Delta\rho(d, d^* \mid x)$ is the expected effect on the response of an intervention $d \to d^*$, *averaging over* unobserved U, *conditional on* observed $X = x$. The unobserved U is not "held constant," but is averaged over; the observed covariates X are not "held constant," but are conditioned on. These distinctions are important: averaging and conditioning are stochastic operations, whereas "holding constant" is a counterfactual operation meaningful only for interventions. Formally, elements $d_j \to d_j^*$ of $d \to d^*$ such that $d_j^* - d_j = 0$ are *held constant.*

In contrast, differences of the form $\rho(d \mid x) - \rho(d \mid x^*)$ have no necessary structural interpretation. Instead they inform us only as to how the predicted response at d varies with different covariate outcomes $X = x$ and $X = x^*$. This is analogous to the fact discussed by CW that in linear regression with conditioning instruments, some coefficients can be meaningfully interpreted as ceteris paribus effects (those associated with D), whereas others may have only a predictive interpretation (those associated with X).

Altonji and Matzkin (2005), Hoderlein (2005, 2007) and Hoderlein and Mammen (2007) pay particular attention to average marginal effects. Here, we consider the *average marginal effect on Y of D_j at d given $X = x$*, defined as

$$\xi_j(d \mid x) := \int \mathsf{D}_j r\,(d, z, u)\, dG(u \mid x),$$

where $\mathsf{D}_j := (\partial / \partial d_j)$, provided the indicated derivative and integral exist. We also call $\xi_j(d \mid x)$ a "covariate-conditioned average marginal effect," or just an "average marginal effect." This is related to (indeed underlies) the average derivatives of Stoker (1986) and Powell, Stock, and Stoker (1989). It is a weighted average of the unobservable marginal effect $\mathsf{D}_j r(d, z, u)$, averaging over unobserved causes, given observed covariates. $\xi_j(d \mid X)$ is the mean squared error-optimal predictor of $\mathsf{D}_j r\,(d, Z, U)$ given X.

When the limit exists, the derivative of ρ with respect to d_j is given by

$$\mathsf{D}_j \rho(d \mid x) := \lim_{\epsilon \to 0}\ \Delta\rho(d, d + \epsilon\ \iota_j \mid x)\ /\ \epsilon,$$

where ι_j is the $k \times 1$ unit vector with unity in the jth position. Typically, the validity of an interchange of derivative and integral is simply assumed, so that

$$\mathsf{D}_j \rho(d \mid x) = \xi_j(d \mid x).$$

The next condition makes explicit general conditions ensuring the existence of derivatives of interest and justifying the interchange of derivative and integral. Further, when ρ and μ are fully identified, this also ensures full identification of $\xi_j(d \mid x)$ as $\mathsf{D}_j \mu(d, x)$. We now let $d_{(j)}$ be the $(k-1) \times 1$ sub-vector of d containing all but d_j, $j \in \{1, \ldots, k\}$.

Assumption A.3 For given $(d, x) \in$ supp (D, X), suppose the function $u \to r(d, z, u)$ is integrable with respect to $G(\cdot \mid x)$, that is, $\int r(d, z, u)\, dG(u \mid x) < \infty$, and suppose that for the given $(d_{(j)}, z)$, $(d_j, z) \to \mathsf{D}_j r\,(d, z, u)$ exists on $C_j \times$ supp $(U \mid z)$, where C_j is a convex compact neighborhood of the given d_j, and supp $(U \mid z)$ is the support of U given $Z = z$. Suppose further that for the given $(d_{(j)}, z)$ and for each u in supp $(U \mid z)$,

$$\sup_{d_j \in C_j} \mid \mathsf{D}_j r\,(d, z, u) \mid \leq \zeta(d_{(j)}, z, u),$$

where ζ is a measurable function such that $E(\zeta(D_{(j)}, Z, U)) < \infty$.

When A.3 holds for r and $\mathsf{D}_j r$, we say "$\mathsf{D}_j r\,(d, z, u)$ is dominated on C_j by an integrable function." The next identification result is a continuation of Theorem 2.2:

Theorem 2.2 *Suppose that the conditions of Theorem 2.2(i.a) hold. (ii.a) If Assumption A.3 also holds, then the functions $d_j \to \rho(d \mid x)$ and $d_j \to \mu(d, x)$ are differentiable on C_j, and $D_j\rho(d \mid x) = \xi_j(d \mid x) = D_j\mu(d, x) = \int D_j r(d, z, u)\, dG(u \mid d, x)$. (ii.b) If Assumption A.1(iii) also holds, then $\xi_j(d \mid x)$ and $D_j\mu(d, x)$ are fully identified.* ■

Thus, to consistently estimate $\xi_j(d \mid x)$, it suffices to consistently estimate $\mathsf{D}_j\mu(d, x)$. If A.3 holds for all $(d, x) \in$ supp (D, X), then ξ_j and $D_j\mu$ are fully identified.

3 Identification of General Effect Measures

Interest also attaches to effects of interventions on aspects of the conditional response distribution other than the mean. Heckman, Smith, and Clements (1997) draw attention to this issue in the context of programme evaluation. Imbens and Newey (2003) discuss a variety of such effects. For wage determination, Firpo, Fortin, and Lemieux (2005) study effects of binary treatments on aspects of the unconditional response distribution, such as the variance, median, or density. Here we discuss identification of structural effects for general aspects of the conditional response distribution, extending the scope of the literature just cited. We consider moment-based effects, where moments may be either explicitly or implicitly defined, as well as effects arising from optimizing behavior.

3.1 Three Ways to Define General Effects

One approach to defining effects uses the *covariate-conditioned counterfactual moment*

$$\rho_0(d \mid x) := \tau_0(\rho_1(d \mid x), \rho_2(d \mid x), \ldots),$$

where τ_0 is a known function, and for known scalar-valued functions τ_k,

$$\rho_k(d \mid x) := \int \tau_k(r(d,z,u))\, dG(u \mid x), \qquad k = 1, 2, \ldots \ .$$

The *moment effect on Y of the intervention $d \to d^*$ to D given $X = x$* is

$$\Delta\rho_0(d, d^* \mid x) := \rho_0(d^* \mid x) - \rho_0(d \mid x),$$

and the *marginal moment effect on Y of D_j given $X = x$* is $\mathsf{D}_j \rho_0(d \mid x)$.

For example, let $\tau_1(r) = 1[r \leq y]$ for $y \in \mathbb{R}$ (cf. Imbens, 2004, p.9), and let $\tau_0(\rho_1) = \rho_1$. Then effects on the conditional response distribution are defined from the counterfactual conditional distribution function

$$\rho_0(d \mid x) = \int 1[r(d,z,u) \leq y]\, dG(u \mid x).$$

Or let $\tau_1(r) = r$, $\tau_2(r) = r^2$, and put $\rho_0(d \mid x) = \rho_2(d \mid x) - \rho_1(d \mid x)^2$. This defines the covariate-conditioned counterfactual variance, yielding conditional variance effects.

When conditional exogeneity holds, the counterfactual moment function and the corresponding effects are stochastically identified.

Theorem 3.1 *Suppose Assumption A.1(i) holds. For $k = 1, 2, \ldots$, let $\tau_k : \mathbb{R} \to \mathbb{R}$ be a known measurable function such that $E(\tau_k(Y)) < \infty$. (i) Then $\mu_k(D, X) := E(\tau_k(Y) \mid D, X)$ exists and is finite, $k = 1, 2, \ldots$. (ii) If A.1(ii) also holds, then for each (d, x) in supp(D, X)*

$$\mu_k(d, x) = \int \tau_k(r(d,z,u))\ dG(u \mid d, x), \qquad k = 1, 2, \ldots \ .$$

If $\tau_0 : R^\infty \to R$ is a known measurable function, then the function μ_0 defined by $\mu_0(d, x) := \tau_0(\mu_1(d, x), \mu_2(d, x), \ldots)$ is also measurable. (iii) If A.2 also holds, then

$$\rho_k(d \mid x) = \int \tau_k(r(d,z,u))\ dG(u \mid x)$$

exists and is finite for each (d, x) in supp(D, X), and $\rho_k = \mu_k$, $k = 1, 2, \ldots$; the function ρ_0 defined by $\rho_0(d \mid x) = \tau_0(\rho_1(d \mid x), \rho_2(d \mid x), \ldots)$ is measurable; and $\rho_0 = \mu_0$. (iv) If A.1(iii) also holds, then ρ_0 and μ_0 are fully identified. ■

General effects also arise from the covariate-conditioned counterfactual optimizer

$$\rho_0(d \mid x) := \arg\max_m \int \tau(r(d,z,u),m)\, dG(u \mid x),$$

where $\tau : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}$ is known, so that $\rho_0(d \mid x)$ is a $\lambda \times 1$ vector of aspects of the counterfactual conditional distribution. For example, effects on the conditional α-quantiles of the response arise from

$$\tau(r,m) = -\left| r - m \right| (\alpha 1[r \geq m] + (1-\alpha)1[r < m]).$$

Now $\rho_0(d \mid x)$ defines the covariate-conditioned counterfactual α-quantile function, a conditional analog of the "quantile structural function" of Imbens and Newey (2003). The associated effect is the covariate-conditioned analog of the quantile treatment effect of Lehmann (1974) and Abadie, Angrist, and Imbens (2002).

When m is a vector and $\tau(r,m)$ defines a quasi-log-likelihood function, this method focuses attention simultaneously on multiple aspects of the counterfactual conditional response distribution, such as location, scale, or quantiles. Taking $\tau(r,m)$ to define an agent's utility function, as in Elliott and Lieli (2005), Skouras (2007), or Lieli and White (in press), yields distributional aspects, $\rho_0(d \mid x)$, of the counterfactual response that determine optimal decisions. When the agent controls D, these aspects make possible the determination of *expected utility-optimal settings*

$$d^*(x) = \arg \max_{d \in \, \mathrm{supp}(D|X=x)} \varsigma_\tau(d \mid x;\ \rho_0(d \mid x)),$$

where ς_τ and ρ_0 are as defined below. Such settings lie at the heart of optimizing behavior under uncertainty.

Theorem 3.2 *Suppose Assumption A.1(i) holds.* (*i*) *For $\lambda \in N$, let $\tau : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}$ be a known measurable function such that $E(\tau(Y,m)) < \infty$ for each m in $\mathbb{R}^\lambda$. Then for each m in $\mathbb{R}^\lambda$, $\varphi_\tau(D,X;m) := E(\tau(Y,m) \mid D,X)$ exists and is finite.* (*ii*) *Suppose A.1(ii) also holds. Then for each (d,x,m) in supp$(D,X) \times \mathbb{R}^\lambda$*

$$\varphi_\tau(d,x;\ m) = \int \tau(r(d,z,u),m)\ dG(u \mid d,x)$$

exists and is finite. Further, let $\tau, r,$ and $(d,x) \to G(\cdot \mid d,x)$ be such that $\varphi_\tau(d,x;m)$ defines a continuous real-valued function on supp$(D,X) \times \mathbb{R}^\lambda$, and let $M\colon$ supp$(D,X) \to \mathbb{R}^\lambda$ be a non-empty and compact-valued continuous correspondence. Then for each (d,x) in supp (D,X) the correspondence

$$\mu_0(d,x) = \arg \max_{m \in M(d,x)} \varphi_\tau(d,x;m)$$

is non-empty, compact-valued, and upper hemi-continuous. (iii) If A.2 also holds, then

$$\varsigma_\tau(d \mid x;\ m) := \int \tau(r(d,z,u),m)\ dG(u \mid x)$$

defines a continuous real-valued function on supp $(D,X) \times \mathbb{R}^\lambda$ such that for each (d,x,m) in supp$(D,X) \times \mathbb{R}^\lambda$ we have $\varsigma_\tau(d \mid x;\ m) = \varphi_\tau(d,x;m)$; the correspondence

$$\rho_0(d \mid x) = \arg \max_{m \in M(d,x)} \varsigma_\tau(d \mid x;\ m)$$

is non-empty, compact-valued, and upper hemi-continuous; and $\rho_0 = \mu_0$. (iv) If A.1(iii) also holds then ρ_0 and μ_0 are fully identified. ■

A third way to define effects uses implicitly defined moments $\rho_0(d,x)$ such that

$$\int \tau(r(d,z,u), \rho_0(d,x))\, dG(u \mid x) = 0,$$

where $\tau : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}^\lambda$ is known. This method has not been previously studied to the best of our knowledge, although it contains many instances of the moment or optimizer approaches as special cases. For example, this $\rho_0(d,x)$ can represent first order conditions defining the interior optimizer of some objective function. The implicit moment approach generalizes and complements the optimizer approach in the same way that method of moments estimation generalizes and complements maximum likelihood estimation.

Theorem 3.3 *Suppose Assumption A.1(i) holds. (i) For $\lambda \in N$, let $\tau : \mathbb{R} \times \mathbb{R}^\lambda \to \mathbb{R}^\lambda$ be a measurable function such that $E(\tau(Y,m)) < \infty$ for each $m \in M \subset \mathbb{R}^\lambda$. Then for each $m \in M$, $\psi_\tau(D,X;m) := E(\tau(r(D,Z,u),m) \mid D,X)$ exists and is finite. (ii) Suppose*

A.1(ii) also holds. Then for each (d, x, m) in supp$(D, X) \times M$

$$\psi_\tau(d, x; m) = \int \tau(r(d, z, u), m)\ dG(u \mid d, x)$$

exists and is finite. Further, let τ, r, and $(d, x) \to G(\cdot \mid d, x)$ be such that for each (d, x, m) in supp$(D, X) \times M$, ψ_τ is differentiable on a neighborhood of (d, x, m), the $\lambda \times \lambda$ matrix $\nabla_m \psi_\tau(d, x; m)$ is non-singular, and $\psi_\tau(d, x; m) = 0$. Then there exists a unique function μ_0 such that for each $(d, x) \in$ supp(D, X), μ_0 is differentiable at (d, x), and

$$\int \tau(r(d, z, u), \mu_0(d, x))\ dG(u \mid d, x) = 0.$$

(iii) If A.2 also holds, then there exists a unique function ρ_0 such that for each $(d, x) \in$ supp(D, X), ρ_0 is differentiable at (d, x);

$$\int \tau(r(d, z, u), \rho_0(d \mid x))\ dG(u \mid x) = 0;$$

and $\rho_0 = \mu_0$. (iv) If A.1(iii) also holds then ρ_0 and μ_0 are fully identified. ■

In each case, for admissible (d, d^*, x), the *$\rho-$effect of the intervention $d \to d^*$ to D given $X = x$* is

$$\Delta\rho(d, d^* \mid x) := \rho(d^* \mid x) - \rho(d \mid x).$$

Our results identify these in terms of the corresponding μ as

$$\Delta\rho(d, d^* \mid x) = \Delta\mu(d, d^*, x) := \mu(d^*, x) - \mu(d, x).$$

The value x need not be factual, so there is an allowed counterfactual aspect to conditioning, although this is not a structural aspect. Comparing the expected effects of an intervention $d \to d^*$ for different values x and x^* gives an *average ρ-effect difference* $\Delta\rho(d, d^* \mid x^*) - \Delta\rho(d, d^* \mid x)$. This measures the impact on effect *expectations* of a "change" in the covariates.

In Section 2.2, we studied average marginal effects. We defer treating general marginal effects to Section 4, where we study weaker conditions for identification.

3.2 Unconditional Effects

Covariate-conditioned effects can be used to construct unconditional effects. For given (d, d^*), let F be a distribution supported on a subset of $\text{supp}(X)$ on which $\Delta\rho(d, d^* \mid x)$ is defined, namely, $\text{supp}(X \mid D = d^*) \cap \text{supp}(X \mid D = d)$. This ensures the analog of the common support assumption (cf. Imbens and Newey, 2003). The unconditional$(-F)$ ρ-effect mean is then given by

$$m_1(d, d^*; \Delta\rho, F) := \int \Delta\rho(d, d^* \mid x)\, dF(x).$$

When $\Delta\rho$ is fully identified, then so is m_1. For example, consider a single binary treatment, with $\Delta\rho(0, 1 \mid x)$ the covariate-conditioned average effect of treatment. Let $F = F_1$, the covariate distribution for the treated observations $(D = 1)$, assuming common support. Then $m_1(0, 1; \Delta\rho, F_1)$ is the average effect of treatment on the treated (e.g., Rubin, 1974).

Similarly, the unconditional$(-F)$ marginal ρ-effect mean is

$$m_1(d; \mathsf{D}_j\rho, F) := \int \mathsf{D}_j\rho(d \mid x)\, dF(x).$$

Other descriptors of the distribution of $\rho-$effects can be straightforwardly defined using other unconditional$(-F)$ ρ-effect moments, such as the kth moments,

$$\begin{aligned} m_k(d, d^*; \Delta\rho, F) &\ : \ = \int \Delta\rho(d, d^* \mid x)^k\, dF(x) \quad \text{or} \\ m_k(d; \mathsf{D}_j\rho, F) &\ : \ = \int \mathsf{D}_j\rho(d \mid x)^k\, dF(x). \end{aligned}$$

3.3 Implications for Estimation

When ρ is fully identified, it suffices to estimate the corresponding stochastic aspect μ, as effect estimates follow by taking suitable differences or derivatives. We now briefly sketch the construction of estimators.

First, consider the covariate-conditioned optimizer of Theorem 3.2,

$$\mu_0(d, x) := \arg\max_m \int \tau(r(d, z, u), m)\ dG(u \mid d, x).$$

This implies

$$\mu_0 = \arg\max_{\mu} \int \tau(r(d,z,u), \mu(d,x)) \; dG(u \mid d,x) \; dH(d,x).$$

To estimate μ_0, parameterize μ by specifying a function of parameters θ, defined by $m(d,x,\theta)$, say, such that $m(d,x,\theta^*) = \mu_0(d,x)$ for all (d,x) in supp (D,X). Then θ^* solves

$$\max_{\theta \in \Theta} \int \tau(r(d,z,u), m(d,x,\theta)) \; dG(u \mid d,x) \; dH(d,x),$$

where Θ is an appropriate finite or infinite dimensional parameter space.

Given a sample of n observations (Y_i, D_i, X_i), an estimator $\hat{\theta}_n$ of θ^* solves

$$\max_{\theta \in \Theta_n} n^{-1} \sum_{i=1}^{n} \tau(Y_i, m(D_{i,}X_i, \theta)),$$

where $\{\Theta_n\}$ is a suitable sequence of subsets of Θ. If $\Theta_n = \Theta$ and Θ is finite dimensional, the method is parametric. Semi-parametric and nonparametric methods are handled by letting Θ be infinite dimensional. For example, one may apply the method of sieves (Grenander, 1981; Chen, 2005).

Next, consider the covariate-conditioned implicit moment μ_0 such that

$$\int \tau(r(d,z,u), \mu_0(d,x)) \; dG(u \mid d,x) = 0.$$

As above, parameterize μ_0, specifying a parameter space Θ and a function $\theta \rightarrow m(d,x,\theta)$, such that $m(d,x,\theta^*) = \mu_0(d,x)$ for all (d,x) in supp (D,X). Then θ^* satisfies

$$\int \tau(r(d,z,u), m(d,x,\theta^*)) \; dG(u \mid d,x) \; dH(d,x) = 0.$$

That is, θ^* solves the implicit moment conditions $E[\tau(Y, m(D,X,\theta^*))] = 0$. To estimate θ^*, apply parametric, semi-parametric, or nonparametric versions of the method of moments. For example, Hansen (1982) and Ai and Chen (2003) describe the properties of such estimators under general conditions. One can also apply Owen's (1988, 2001) empirical likelihood methods (see also, e.g., Schennach (2007) or Ragusa (2005)).

Kernel methods apply straightforwardly to nonparametric estimation of explicit moments. See, e.g., Pagan and Ullah (1999), Li and Racine (2007). Li, Lu, and Ullah

(2003) and Schennach, White, and Chalak (2007) give results for estimating derivatives of conditional expectations.

The results of Section 3.1 thus ensure structural interpretations, based on the use of covariates, for fully identified aspects of the response distribution estimated by parametric, semi-parametric, or nonparametric extremum and method of moments procedures. As this covers most procedures commonly used in econometrics, this ensures that each of these methods can exploit covariates to deliver structural content. In particular, under full identification, differences and derivatives of estimators $\hat{\mu}_n = m(\cdot, \cdot, \hat{\theta}_n)$ with respect to elements of causes of interest D have structural meaning. In line with our discussion of Section 2.2, differences and derivatives with respect to elements of the covariates X have expectational or predictive content, but no necessary structural meaning.

Estimators of unconditional effects are given by $m_1(d, d^*; \Delta\hat{\mu}_n, \hat{F}_n)$ and $m_1(d; \mathsf{D}_j\hat{\mu}_n, \hat{F}_n)$, for example, where $\hat{F}_n$ is an estimator of F, such as an empirical distribution or a smoothed empirical distribution.

4 Identification Without Conditional Exogeneity

We now study identification without conditional exogeneity. This yields conditions ensuring identification at specific values (d, x), i.e., locally. In some cases, we obtain necessary and sufficient conditions. We also obtain "near identification" results.

4.1 Explicit Moment Effects

We first consider the relationship between ρ_k and μ_k of Theorem 3.1 without A.2.

Theorem 4.1 *Suppose that A.1(i) holds and let*

$$s(d, x, u) := 1 - dG(u \mid x) \,/\, dG(u \mid d, x) = 1 - dG(d \mid x) \;/\; dG(d \mid u, x).$$

(i) Then for all $(d, x) \in$ *supp* (D, X), $\int s(d, x, u)\ dG(u \mid d, x)) = 0$; *(ii) Further, let A.1(ii) and the remaining conditions of Theorem 3.1(i) hold, and suppose that* $E(s(D, X, U)^2) < \infty$ *and* $E(\tau_k(Y)^2) < \infty$, $k = 1, 2, \ldots$. *Then for all* $(d, x) \in$ *supp* (D, X), $\rho_k(d \mid x)$ *as*

defined in Theorem 3.1(ii) exists and is finite, and

$$\mu_k(d,x) = \rho_k(d \mid x) + \gamma_k(d,x),$$

where

$$\gamma_k(d,x) := \int \tau_k(r(d,z,u))\ s(d,x,u)\, dG(u \mid d,x), \qquad k = 1, 2, \ldots \quad .$$

(iii) Further, for all $(d,x) \in$ *supp* (D,X),

$$|\gamma_k(d,x)| \leq \sigma(d,x;\tau_k)\ \sigma(d,x;s), \qquad k = 1, 2, \ldots,$$

where $\sigma(d,x;\tau_k) := [$*var* $(\tau_k(Y) \mid (D,X) = (d,x))]^{1/2}$ *and* $\sigma(d,x;s) := [$*var* $(s(D,X,U) \mid (D,X) = (d,x))]^{1/2}$. ■

The *discrepancy score* $s(d,x,u)$ measures the relative departure from conditional exogeneity at (d,x,u). By (i), the discrepancy score has conditional mean zero.

By (ii), $\mu_k(d,x)$ differs from $\rho_k(d \mid x)$ by the *moment discrepancy* $\gamma_k(d,x)$, which, given (i), is the conditional covariance of $\tau_k(Y)$ and $s(D,X,U)$. Thus, $\gamma_k(d,x) = 0$ is necessary and sufficient for stochastic identification of $\rho_k(d \mid x)$. This is a *local identification* result, specific to a particular (d,x). Conditional exogeneity is sufficient for this, but not necessary. It suffices that $s(d,x,u) = 0$ for all $u \in \text{supp}\ (U \mid (D,X) = (d,x))$. Chesher (2003, 2005) and Matzkin (2004) give related results involving local identification under some monotonicity assumptions.

The Cauchy-Schwarz inequality gives (iii), bounding the moment discrepancy and establishing a form of continuity with respect to (i) local dependence of $\tau_k(Y)$ on unobservables, measured by $\sigma(d,x;\tau_k)$; and (ii) departures from local conditional exogeneity, measured by $\sigma(d,x;s)$. If either is small, then so is the moment discrepancy. Theorem 4.1(iii) is thus a near identification result. The bound is best possible, as equality holds when $|\tau_k(r(d,z,u))| = |s(d,x,u)|$ for given (d,x) and all u in supp$(U \mid (D,X) = (d,x))$. Similar results follow by applying the Hölder inequality.

Theorem 3.1 treats $\mu_0(d,x) = \tau_0(\mu_1(d,x),\ \mu_2(d,x),\ldots)$ and $\rho_0(d,x) = \tau_0(\rho_1(d,x),$

$\rho_2(d,x),\ldots)$. The *general moment discrepancy* is

$$\gamma_0(d,x) := \mu_0(d,x) - \rho_0(d,x).$$

If τ_0 is affine in its arguments (e.g., the covariate-conditioned average response),

$$\gamma_0(d,x) = \tau_0(\gamma_1(d,x), \gamma_2(d,x), \ldots).$$

Theorem 4.1(ii) then implies that the "apparent effect" $\Delta\mu_0(d,d^*,x)$ is contaminated by the *effect discrepancy*

$$\Delta\gamma_0(d,d^*,x) = \tau_0(\Delta\gamma_1(d,d^*,x), \Delta\gamma_2(d,d^*,x), \ldots),$$

where $\Delta\gamma_k(d,d^*,x) := \gamma_k(d^*,x) - \gamma_k(d,x), k = 1,2,\ldots$.

Even if τ_0 is not affine, γ_0 depends globally and smoothly on the γ_k's, under plausible conditions. For brevity, let τ_0 depend continuously on $\mu := (\mu_1, \ldots, \mu_\kappa)'$ taking values in a compact set, K. Then

$$\tau_0(\mu) = \sum_{i=1}^{\infty} a_i \text{cos }(\mu'\theta_i) + \sum_{i=1}^{\infty} b_i \sin(\mu'\theta_i),$$

gives the Fourier series representation, where a_i's and b_i's are Fourier coefficients, θ_i's are appropriate multi-frequencies, and equality is in the sense of uniform convergence. Then

$$\tau_0(\mu) - \tau_0(\rho) = [\sum_{i=1}^{\infty} a_i \cos(\mu'\theta_i) - \sum_{i=1}^{\infty} a_i \cos(\rho'\theta_i)] + [\sum_{i=1}^{\infty} b_i \sin(\mu'\theta_i) - \sum_{i=1}^{\infty} b_i \sin(\rho'\theta_i)]$$

for any $\mu, \rho \in K$. Standard trigonometric identities give

$$\begin{aligned} \cos(u) - \cos(v) &= 2\sin(u)\cos([v-u]/2)\sin([v-u]/2) + 2\cos(u)\sin^2([v-u]/2) \\ \sin(u) - \sin(v) &= -2\cos(u)\cos([v-u]/2)\sin([v-u]/2) + 2\sin(u)\sin^2([v-u]/2). \end{aligned}$$

Letting $\gamma := \mu - \rho$ and substituting into $\tau_0(\mu) - \tau_0(\rho)$ then gives

$$\begin{aligned}\gamma_0(\mu, \gamma) \;\; &= \;\; 2\sum_{i=1}^{\infty}[a_i \sin(\mu'\theta_i) - b_i \cos(\mu'\theta_i)] \cos(\gamma'\theta_i/2) \sin(\gamma'\theta_i/2) \\ & \quad + 2\sum_{i=1}^{\infty}[a_i \cos(\mu'\theta_i) + b_i \sin(\mu'\theta_i)] \sin^2(\gamma'\theta_i/2).\end{aligned}$$

The general moment discrepancy γ_0 thus depends globally and smoothly on γ. Theorem 4.1(iii) then ensures that the effect discrepancy $\Delta\gamma_0$ inherits continuity with respect to both conditional dependence and local dependence of the response on unobservables.

Thus, neglecting to proxy unobservables that have minor relevance for determining either the response or the cause of interest leads to correspondingly minor distortions in the apparent effect. Priority should therefore be given to including proxies for those unobservables most relevant to determining the response and causes of interest.

Marginal effects are similarly affected when A.2 fails. We add further structure, gaining analytic convenience without losing much generality. We now require that possible values for U do not depend on the realization of D, though they may depend on that of X. This permits conditional dependence, as the probabilities associated with these possible values can depend on the realization of D.

Recall that a σ-finite measure η is absolutely continuous with respect to a σ-finite measure ν, written $\eta \ll \nu$, if $\eta(B) = 0$ for every measurable set B such that $\nu(B) = 0$. If $\eta \ll \nu$, we say ν *dominates* η and call ν a "dominating measure." The Radon-Nikodym theorem states that if $\eta \ll \nu$, then there exists a positive measurable function $f = d\eta/d\nu$, the *Radon-Nikodym density*, such that $\eta(A) = \int_A f \; d\nu$ for every measurable set A.

Assumption A.4 For each $x \in$ supp X, there exists a σ-finite measure $\nu(\,\cdot \mid x)$ such that (i) for each $(d, x) \in \text{supp}(D, X)$, the measure $G(B \mid d, x) = \int_B dG(u \mid d, x)$ is absolutely continuous with respect to $\nu(\,\cdot \mid x)$; (ii) for each $x \in$ supp (X), the measure $G(B \mid x) = \int_B dG(u \mid x)$ is absolutely continuous with respect to $\nu(\,\cdot \mid x)$.

By Radon-Nikodym, there exist conditional densities, say $g(u \mid d, x)$ and $g(u \mid x)$ such that $dG(u \mid d, x) = g(u \mid d, x) \; d\nu(u \mid x)$ and $dG(u \mid x) = g(u \mid x) \; d\nu(u \mid x)$. Conditional dependence arises whenever $g(u \mid d, x)$ depends non-trivially on d.

Next, we impose differentiability and domination conditions.

Assumption A.5 For given j, $\mathsf{D}_j g\,(u \mid d, x)$ is dominated on C_j by a function integrable with respect to $\nu(\,\cdot \mid x)$ at (d, x).

We also impose the analog of A.3:

Assumption A.6 For given j and $k = 1, 2, \ldots$, $\mathsf{D}_j[(\tau_k \circ r)g](d, x, u)$ is dominated on C_j by a function integrable with respect to $\nu(\,\cdot \mid x)$ at (d, x).

The differentiability of g implicit in A.5 and differentiability of $\tau_k \circ r$ with respect to d_j implicit in A.6 ensure existence of the product derivative $\mathsf{D}_j[(\tau_k \circ r)g]$.

Theorem 4.2 *Suppose Assumptions A.1(i) and A.4 hold and let* $s(d, x, u) := 1 - g(u \mid x) \;/\; g(u \mid d, x) = 1 - g(d \mid x)/g(d \mid u, x)$ *(i.a) Then for all* $(d, x) \in$ *supp* (D, X), $\int s(d, x, u)\, g(u \mid d, x)\; d\nu(u \mid x) = 0$. *(i.b) If A.5 also holds, then for the given* (d, x)

$$\int \mathsf{D}_j \,\log g\,(u \mid d, x)\, g(u \mid d, x)\; d\nu(u \mid x) = 0.$$

(ii) Further, let A.1(ii), A.6, and the remaining conditions of Theorem 3.1(i) hold. Then the functions $d_j \to \mu_k(d, x)$, $k = 1, 2, \ldots$, *are differentiable on* C_j *and*

$$\mathsf{D}_j \mu_k(d, x) = \int \mathsf{D}_j \tau_k(r(d, z, u))\;\, g(u \mid d, x)\; d\nu(u \mid x)$$

$$+ \int \tau_k(r(d, z, u))\, \mathsf{D}_j \,\log g(u \mid d, x)\, g(u \mid d, x)\; d\nu(u \mid x).$$

(iii) If, in addition for $k = 1, 2, \ldots$ $E([D_j \tau_k(r(D, Z, U))]^2) < \infty$ *and* $E(s(D, X, U)^2) < \infty$*, then for* $k = 1, 2, \ldots$

$$\begin{array}{lcl}
\mathsf{D}_j \mu_k(d, x) & = & \xi_{k,j}(d \mid x) + \delta_{1,k,j}(d, x) + \delta_{2,k,j}(d, x), \qquad\qquad \textit{where} \\
\xi_{k,j}(d \mid x) & := & \displaystyle\int \mathsf{D}_j \tau_k(r(d, z, u))\, g(u \mid x)\; d\nu(u \mid x) \\
\delta_{1,k,j}(d, x) & := & \displaystyle\int \mathsf{D}_j \tau_k(r(d, z, u))\, s(d, x, u)\, g(u \mid d, x)\; d\nu(u \mid x) \\
\delta_{2,k,j}(d, x) & := & \displaystyle\int \tau_k(r(d, z, u))\, \mathsf{D}_j \,\log g(u \mid d, x)\; g(u \mid d, x)\; d\nu(u \mid x).
\end{array}$$

(iv)(a) Letting $\sigma(d, x; D_j(\tau_k \circ r)) := [\;\textit{var}\;(D_j \tau_k(r(D, Z, U)) \mid (D, X) \;=\; (d, x))]^{1/2}$,

$$\mid \delta_{1,k,j}(d, x) \mid \leq \sigma(d, x; \mathsf{D}_j(\tau_k \circ r))\, \sigma(d, x; s).$$

(b) If in addition $E([D_j \log g(U \mid D, X)]^2) < \infty$ *and* $E(\tau_k(Y)^2) < \infty$*,* $k = 1, 2, \ldots$*, then*

$$|\delta_{2,k,j}(d,x)| \leq \sigma(d,x;\tau_k)\, \sigma(d,x;\mathsf{D}_j \log g_d)$$

where $\sigma(d,x;D_j \log g_d) := [\int \{D_j \log g(u \mid d,x)\}^2\, g(u \mid d,x)\; d\nu(u \mid x)]^{1/2}$. ■

Thus, $\mathsf{D}_j\mu_k$ is a contaminated version of $\xi_{k,j}$, the covariate-conditioned average marginal τ_k moment effect. The effect discrepancy is $\delta_{k,j} := \delta_{1,k,j} + \delta_{2,k,j}$. Conditional exogeneity is sufficient but not necessary for this to vanish. The effect discrepancy component $\delta_{1,k,j}(d,x)$ vanishes if the discrepancy score $s(d,x,u)$ vanishes for all u. The effect discrepancy component $\delta_{2,k,j}(d,x)$ vanishes if the *marginal discrepancy score* D_j log $g(u \mid d,x)$ vanishes for all u. Result (iv) bounds the moment discrepancies. We thus have local identification and near identification results analogous to those of Theorem 4.1.

Results relating to $\mathsf{D}_j\mu_0$ to $\mathsf{D}_j\rho_0$ follow straightforwardly. For brevity, let τ_0 depend on a $\kappa \times 1$ vector, and write $\mu := (\mu_1, \ldots, \mu_\kappa)'$ and $\rho := (\rho_1, \ldots, \rho_\kappa)'$. The chain rule gives $\mathsf{D}_j\mu_0 = \nabla'\tau_0(\mu)(\mathsf{D}_j\mu), \mathsf{D}_j\rho_0 = \nabla'\tau_0(\rho)(\mathsf{D}_j\rho)$,where $\nabla\tau_0$ is the $\kappa \times 1$ gradient vector of τ_0 with respect to its arguments, and $\mathsf{D}_j\mu$ and $\mathsf{D}_j\rho$ are $\kappa \times 1$ vectors (D_j operates element by element). Adding and subtracting appropriately gives

$$\mathsf{D}_j\mu_0 - \mathsf{D}_j\rho_0 = \nabla'\tau_0(\mu)[\mathsf{D}_j\mu - \mathsf{D}_j\rho]$$
$$+[\nabla'\tau_0(\mu) - \nabla'\tau_0(\rho)]\,(\mathsf{D}_j\mu) - [\nabla'\tau_0(\mu) - \nabla'\tau_0(\rho)][\mathsf{D}_j\mu - \mathsf{D}_j\rho]$$

$$= \nabla'\tau_0(\mu)\delta_j + \nabla'\gamma_0(\mu,\gamma)(\mathsf{D}_j\mu) - \nabla'\gamma_0(\mu,\gamma)\delta_j,$$
$$=: \delta_0(\mu,\gamma,\delta_j),$$

where δ_j is the vector with elements $\delta_{k,j}$ and $\nabla'\gamma_0(\mu,\gamma) = \nabla'\tau_0(\mu) - \nabla'\tau_0(\rho)$ holds with $\gamma := \mu - \rho$ and smoothness assumptions on τ_0 sufficient for the Fourier series approximation above to hold in a suitable Sobolev norm. The *marginal general moment effect discrepancy* $\delta_0(\mu,\gamma,\delta_j)$ can vanish for specific values of (d,x) under special circumstances. It vanishes for all (d,x) under conditional exogeneity.

4.2 Implicit Moment Effects

We subsume optimizer-based effects into the study of implicit moment effects, viewing optimizer-based distributional aspects as implicit moments defined by the first order conditions of the underlying optimization.

The implicit nonlinear definitions of μ_0 and ρ_0 present significant challenges to directly obtaining a tractable representation for the implicit moment discrepancy $\gamma_0 := \mu_0 - \rho_0$, so for brevity we do not treat this here. Nevertheless, implicitly defined moments can often be well approximated by the explicit moments analyzed in Theorems 3.1 and 4.1.

An analysis of marginal effects analogous to Theorem 4.2 is more straightforward. To succinctly state our next result, we now write

$$\int \tau_\mu \; g_d \; d\nu = \int \tau(r(d,z,u), \mu_0(d,x)) \; g(u \mid d,x) \; d\nu(u \mid x).$$

When the integral of Theorem 4.3(ii) below exists, we write

$$\int \tau_\rho \; g \; d\nu = \int \tau(r(d,z,u), \rho_0(d \mid x)) \; g(u \mid x) \; d\nu(u \mid x).$$

When the referenced derivatives exist, we now write $\mathsf{D}_j\mu_0$ as the $\lambda \times 1$ vector containing the derivatives $\mathsf{D}_j\mu_{0,i}(d,x)$, $i = 1,\ldots,\lambda$, and $\mathsf{D}_j\rho_0$ is now the $\lambda \times 1$ vector containing $\mathsf{D}_j\rho_{0,i}(d \mid x)$, $i = 1,\ldots,\lambda$. For $\tau(r,m)$, let $\nabla_r\tau_\rho$ denote the $\lambda \times 1$ vector containing $(\partial/\partial r)\tau_i(r(d,z,u),\, \rho_0(d \mid x))$, $i = 1,\ldots,\lambda$; let $\nabla_r\tau_\mu$ denote the $\lambda \times 1$ vector containing $(\partial/\partial r)\tau_i(r(d,z,u),\, \mu_0(d,x)), i = 1,\ldots,\lambda$; let $\nabla'_m\tau_\mu$ denote the $\lambda \times \lambda$ matrix whose ith row has elements $(\partial/\partial m_j)\tau_i(r(d,z,u),\, \mu_0(d,x))$, $j = 1,\ldots,\lambda$, $i = 1,\ldots,\lambda$; and let $\nabla'_m\tau_\rho$ denote the $\lambda \times \lambda$ matrix whose ith row has elements $(\partial/\partial m_j)\tau_i(r(d,z,u),\, \rho_0(d \mid x))$, $j = 1,\ldots,\lambda$, $i = 1,\ldots,\lambda$. When the integrals and inverses exists, we define $Q_\mu := -\int \nabla'_m\tau_\mu \; g_d \; d\nu$, $Q_\rho := -\int \nabla'_m\tau_\rho \; g \; d\nu$.

Assumption A.7 (i) The elements of $\tau(r(d,z,u),\, \mu_0(d,x)) \; g(u \mid d,x)$ are dominated on C_j by a function integrable with respect to $\nu(\cdot \mid x)$ at (d,x). (ii) The elements of $\tau(r(d,z,u),\, \rho_0(d \mid x))$ are dominated on C_j by a function integrable with respect to $dG(\cdot \mid x)$ at (d,x).

Theorem 4.3. *(i) Suppose the conditions of Theorem 3.3(i) and A.1(ii) hold, that*

$E(s(D,X,U)^2) < \infty$, and that $E(\tau(Y,m)^2) < \infty$ for each $m \in M \subset R^\lambda$. Then for each (d,x,m) in supp $(D,X) \times M$ the conditional expectation

$$\psi_\tau(d,x,m) = \int \tau(r(d,z,u),m)\ dG(u \mid x)$$

exists and is finite. (ii) Further, let τ, r, and $x \to G(\cdot \mid x)$ be such that for each (d,x,m) in supp $(D,X) \times M$, ψ_τ is differentiable on a neighborhood of (d,x,m), the $\lambda \times \lambda$ matrix $\nabla_m \psi_\tau(d,x,m)$ is non-singular, and $\psi_\tau(d,x,m) = 0$. Then there exists a unique function ρ_0 such that for each $(d,x) \in$ supp (D,X), ρ_0 is differentiable at (d,x) and

$$\int \tau(r(d,z,u),\ \rho_0(d \mid x))\ dG(u \mid x) = 0.$$

(iii) If A.4(ii) and A.7(ii) also hold, if τ is differentiable and for the given $(d_{(j)},x)$, $(d_j,u) \to D_j r(d,z,u)$ exists on $C_j \times$ supp $(U \mid z)$, and if Q_ρ exists and is finite and non-singular, then

$$\mathsf{D}_j \rho_0 = Q_\rho^{-1} \int \nabla_r \tau_\rho\ (\mathsf{D}_j r)\ g\ d\nu.$$

(iv) If in addition A.4(i) and A.7(i) hold and if Q_μ exists and is finite and nonsingular, then $D_j \mu_0 = D_j \rho_0 + \delta_{0,j}$, where

$$\begin{aligned}
\delta_{0,j} &:= \delta_{1,j} + \delta_{2,j} + \delta_{3,j} \\
\delta_{1,j} &:= Q_\rho^{-1} \int \nabla_r \tau_\rho (\mathsf{D}_j r)\, s\, g_d\, dv \\
\delta_{2,j} &:= Q_\mu^{-1} \int \tau_\mu\, (\mathsf{D}_j\ \log\, g_d)\, g_d\, dv
\end{aligned}$$

$$\delta_{3,j} := \int (Q_\mu^{-1} \nabla_r\, \tau_\mu - Q_\rho^{-1} \nabla_r\, \tau_\rho)\ \mathsf{D}_j r\, g_d\, dv.$$

(v)(a) Suppose $E[(\triangledown_r \tau(Y, \rho_0(D \mid X))'\ \triangledown_r \tau(Y, \rho_0(D \mid X))\ (D_j r(D,Z,U))^2] < \infty$, define $\tilde{\sigma}(\ \cdot\ ; \triangledown_r \tau\ D_j r) :=\ [\int \triangledown_r \tau'_\rho\ \triangledown_r\ \tau_\rho (D_j r)^2\ g_d\ dv]^{1/2}$, and let $\bar{\lambda}_\rho$ denote the largest eigenvalue of $(Q_\rho^{-1\prime} Q_\rho^{-1})$. Then

$$\|\delta_{1,j}\| := (\delta'_{1j} \delta_{1j})^{1/2} \le \bar{\lambda}_\rho^{1/2}\ \tilde{\sigma}(\ \cdot\ ; \triangledown_r \tau \mathsf{D}_j r)\ \sigma(\ \cdot\ ; s).$$

(b) Suppose $E[\tau(Y, \mu_0(D,X))'\ \tau(Y, \mu_0(D,X))] < \infty$, define $\tilde{\sigma}(\ \cdot\ , \tau_\mu) := [\int \tau'_\mu \tau_\mu\ g_d\ dv]^{1/2}$,

and let $\bar{\lambda}_\mu$ *denote the largest eigenvalue of* $(Q_\mu^{-1\prime} Q_\mu^{-1})$. *Then*

$$\|\delta_{2,j}\| \leq \bar{\lambda}_\mu^{1/2}\ \tilde{\sigma}(\ \cdot\ ;\tau_\mu)\ \sigma(\ \cdot\ ;\mathsf{D}_j \log g_d).$$

(c) Suppose $E\ [\bigtriangledown_r \tau(Y,\mu_0(D,X))'\ \bigtriangledown_r \tau(Y,\mu_0(D,X))] < \infty$ *and* $E[\bigtriangledown_r\ \tau(Y,\rho_0(D \mid X))'\ \bigtriangledown_r\ \tau\ (Y,\rho_0(D \mid X))] < \infty$, *and define* $\tilde{\sigma}(\ \cdot\ ; Q_\mu^{-1} \bigtriangledown_r \tau_\mu - Q_\rho^{-1} \bigtriangledown_r \tau_\rho) := [\int (Q_\mu^{-1} \bigtriangledown_r \tau_\mu - Q_\rho^{-1} \bigtriangledown_r \tau_\rho)'\ (Q_\mu^{-1} \bigtriangledown_r \tau_\mu - Q_\rho^{-1} \bigtriangledown_r \tau_\rho)\ g_d\ dv]^{1/2}$ *and* $\tilde{\sigma}(\ \cdot\ ; D_j r) := [\int (D_j r)^2\ g_d dv]^{1/2}$. *Then*

$$\|\delta_{3j}\| \leq \tilde{\sigma}(\ \cdot\ ; Q_\mu^{-1} \bigtriangledown_r \tau_\mu - Q_\rho^{-1} \bigtriangledown_r \tau_\rho) \times \tilde{\sigma}(\ \cdot\ ; \mathsf{D}_j r). \quad \blacksquare$$

The functions in $(iii) - (v)$ above are implicitly evaluated at (d, x) as specified in A.7.

To interpret the *marginal implicit moment effect discrepancy* $\delta_{0,j}$, note that each of its components vanishes under conditional exogeneity, as then $s = 0$, $\mathsf{D}_j \log g_d = 0$, and $Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho = 0$ (Theorem 3.3). If conditional exogeneity fails but the marginal effect $\mathsf{D}_j r\ (d, z, u)$ is zero for all u in supp $(U \mid d, x)$, the true marginal implicit moment effect vanishes ($\mathsf{D}_j \rho_0(d \mid x) = 0$), but the apparent effect becomes

$$\delta_{0,j} := Q_\mu^{-1} \int \tau_\mu\ (\mathsf{D}_j \log g_d)\ g_d\ d\nu.$$

The important special case $\tau(r, m) = r - m$ offers further insight. Here $Q_\mu = Q_\rho = 1$ and $\nabla_r \tau_\mu = \nabla_r \tau_\rho = 1$. In this case, we have (cf. Theorem 4.2(iii))

$$\delta_{0,j} := \int (\mathsf{D}_j r)\ s\ g_d\ d\nu + \int\ \tau_\mu (\mathsf{D}_j \log g_d)\ g_d\ d\nu.$$

Theorem 4.3 affords considerable opportunity to explore special cases of interest (for example, when $\lambda = 1$, consider the anti-symmetric case in which $\tau(r, m) = -\tau(m, r)$, which applies to the conditional median). For brevity, we leave this aside here.

The first two components of $\delta_{0,j}$ are clearly conditional covariances, given that s and D_j log g_d have conditional mean zero. The third term is generally not a covariance because there is no need for $\mathsf{D}_j r$ or $Q_\mu^{-1} \nabla_r \tau_\mu - Q_\rho^{-1} \nabla_r \tau_\rho$ to have conditional mean zero. Nevertheless, under a hypothesis of no effect, specifically, that $\mathsf{D}_j r$ has conditional mean zero, the third term is again a covariance. Result (v) provides a near identification and continuity result, generalizing that of Theorem 4.2(iv).

5 Summary and Conclusion

We examine the use of covariates to identify and estimate structural effects of multiple causes. These can be binary, categorical, or continuous. For the case of continuous causes, we examine both marginal and non-marginal effects. We analyze effects generally defined, based on explicit or implicit moments, as well as on aspects of the response distribution defined as optimizers (e.g., quantiles). The latter lead to extremum estimators; the former lead to method of moment estimators. As we show, the procedures commonly used in econometrics, for example parametric, semi-parametric, and nonparametric extremum or moment-based methods, can all exploit covariates to estimate well-identified structural effects. These results hold for general structural systems, without imposing linearity, separability, or monotonicity.

We also study the role of conditional exogeneity by examining what happens in its absence. We find that identification generally fails, although it may hold locally. We also obtain near identification results that provide insight into how the various effect measures are impacted by departures from local conditional exogeneity and the sensitivity of the response of interest to unobservables.

There are a variety of directions for further investigation. Of particular interest are analyzing efficiency bounds for extremum and moment-based estimators of effects under conditional exogeneity in nonseparable settings, developing tests for conditional exogeneity, and developing tests for restrictions such as monotonicity or separability in the context of general nonseparable structures.

A Mathematical Appendix

Proof of Proposition 2.1 (i) Given A.1(i) and $E(Y) < \infty$, $E(Y \mid D, X)$ exists and is finite by Billingsley (1979, p.395). (ii) Apply theorem 34.5 of Billingsley (1979). ■

Proof of Theorem 2.2 $(i.a)$ Given A.1(i, ii) and $E(Y) < \infty$, Proposition 2.1 gives $\mu(d, x) = \int r(d, z, u)\ dG(u \mid d, x)$. Assumptions A.1$(ii)$ and A.2 then imply $\int r(d, z, u)\ dG(u \mid d, x) = \int r(d, z, u)\ dG(u \mid x) = \rho(d \mid x)$, ensuring both existence of $\rho(d \mid x)$ and $\rho(d \mid x) = \mu(d, x)$. $(i.b)$ The result follows immediately from the identification definitions given A.1(iii). $(ii.a)$ Assumption A.3 permits application of Bartle (1966, corollary 5.9) establishing the differentiability of ρ and μ (by (i)) and ensuring the validity of an interchange of derivative and integral, giving

$$\begin{aligned}
\mathsf{D}_j\mu(d,x) &= \mathsf{D}_j\rho(d \mid x) \\
&= \int \mathsf{D}_j r\ (d,z,u)\ dG(u \mid x) \qquad (=: \xi_j(d \mid x)) \\
&= \int \mathsf{D}_j r(d,z,u)\ dG(u \mid d,x).
\end{aligned}$$

The first equality holds by $(i.a)$. In the second, A.3 ensures the interchange of derivative and integral, and A.1(ii) ensures the absence of terms involving $(\partial z/\partial d_j)$. The third equality follows by A.2. $(ii.b)$ Immediate, given A.1(iii). ■

Proof of Proposition 2.3 D, U, V, and X exist by Assumption A.1(i). Dawid (1979, lemma 4.3) establishes the if part and Dawid's (1979) lemma 4.2 and the symmetry of conditional independence establish the converse, as Dawid (1979) states. ■

Proof of Theorem 3.1. (i) For given $k = 1, 2, \ldots$, the proof is identical to that of Proposition 2.1(i), *mutatis mutandis* (replacing r with $\tau_k \circ r$). The measurability of μ_0 follows by measurability of compositions of measurable functions. (ii, iii) For given $k = 1, 2, \ldots$, the proof is identical to that of Theorem 2.2$(i.a)$, *mutatis mutandis.* Measurability follows for ρ_0 just as for μ_0. That $\rho_0 = \mu_0$ follows immediately from $\rho_k = \mu_k$, $k = 1, 2, \ldots$. (iv) Immediate, given A.1(iii). ■

Proof of Theorem 3.2. (i) Given $E(\tau(Y,m)) < \infty$, the existence and finiteness of $\varphi_\tau\ (d,x;m)$ follow from Billingsley (1979, p.395). (ii) For each (d,x) in supp (D,X), the existence of the non-empty, compact-valued, upper hemi-continuous correspondence $\mu_0(d,x)$ follows from the Theorem of the Maximum (Berge, 1963) under the stated conditions. (iii) If A.1(ii) and A.2 also hold, then for each $(d,x,m) \in (D,X) \times \mathbb{R}^\lambda$ $\varphi_\tau\ (d,x;m)$ $= \varsigma_\tau\ (d \mid x;m)$. Setting $\rho_0\ (d \mid x) = \mu_0\ (d,x)$ completes the proof. (iv) Immediate, given A.1(iii). ■

Proof of Theorem 3.3. (i) Given $E(\tau(Y,m)) < \infty$, the existence and finiteness of $\psi_\tau\ (d,x,m)$ follow from Billingsley (1979, p. 395). (ii) The existence, uniqueness, and differentiability of μ_0 follow immediately under the given conditions from the implicit function theorem (e.g., Chiang, 1984, pp. 210-211). (iii) If A.1(ii) and A.2 also hold, then for each $(d, x) \in$ supp (D, X)

$$\int \tau(r(d, z, u), \mu_0(d, x)) \; dG(u \mid d, x) = \int \tau(r(d, z, u), \mu_0(d, x)) \; dG(u \mid x).$$

Setting $\rho_0(d \mid x) = \mu_0(d, x)$ completes the proof. (iv) Immediate, given A.1(iii). ■

Proof of Theorem 4.1. Given A.1(i), the densities $dG(u \mid d, x)$ and $dG(u \mid x)$ exist and can be written $dG(u \mid d, x) = dG(d, x, u) \; / \; dG(d, x)$ and $dG(u \mid x) = dG(x, u) \; / \; dG(x)$. Consequently, $dG(u \mid x) \; / \; dG(u \mid d, x) = [dG(x, u) \; / \; dG(x)] \; / \; [dG(d, x, u) \; / \; dG(d, x)] = [dG(d, x) \; / \; dG(x)] \; / \; [dG \; (d, x, u) \; / \; dG(x, u)] = dG(d \mid x) \; / \; dG(d \mid u, x)$. (i) We have $\int s(d, x, u) \; dG(u \mid d, x) = \int \{[dG(u \mid d, x) - dG(u \mid x)] \; / \; dG(u \mid d, x)\} \; dG(u \mid d, x) = \int [dG(u \mid d, x) - dG(u \mid x)] = 0$, given that $dG(u \mid d, x)$ and $dG(u \mid x)$ are each conditional densities. (ii) Given A.1(ii) and the conditions on τ_k, Theorem 3.1(i) gives $\mu_k(d, x) = \int \tau_k(r(d, z, u)) \; dG(u \mid d, x)$. Adding and subtracting appropriately, with A.1(ii) we have

$$\begin{aligned}
\rho_k(d \mid x) \quad &:= \int \tau_k(r(d, z, u)) \; dG(u \mid x) \\
&= \int \tau_k(r(d, z, u)) \; dG(u \mid d, x) + \int \tau_k(r(d, z, u)) \; [dG(u \mid x) - \; dG(u \mid d, x)] \\
&= \mu_k(d, x) + \int \tau_k(r(d, z, u)) \{[dG(u|x) - \; dG(u \mid d, x)] \; / \; dG(u \mid d, x)\} \; dG(u \mid d, x) \\
&= \mu_k(d, x) - \int \tau_k(r(d, z, u)) \; s(d, x, u) \; dG(u \mid d, x) \\
&= \mu_k(d, x) - \gamma_k(d, x).
\end{aligned}$$

The existence of $\mu_k(d, x)$ follows given $E(\tau_k(Y)) < \infty$ and the existence of $\gamma_k(d, x)$ follows from the imposed second moment conditions and the Cauchy-Schwarz inequality. It follows that $\rho_k(d \mid x)$ exists and $\mu_k(d, x) = \rho_k(d \mid x) + \gamma_k(d, x)$. (iii) The result follows immediately from the Cauchy-Schwarz inequality, applied to (ii) and using (i). ■

Proof of Theorem 4.2. $(i.a)$ Assumptions A.1(i) and A.4(i) ensure that for each (d, x) in supp (D, X), $g(u \mid d, x) = dG(u \mid d, x) \; / \; d\nu(u \mid x)$ is a density by the Radon-Nikodym theorem (e.g., Bartle, 1966, theorem 8.9), so $\int g(u \mid d, x) \; d\nu(u \mid x) = 1$. $(i.b)$ Assumption A.5 ensures that the left hand expression above is differentiable with respect to d_j by Bartle (1966, corollary 5.9). Differentiating both sides of this equality with respect to d_j

gives

$$\mathsf{D}_j \int g(u \mid d, x)\ d\nu(u \mid x) = 0.$$

Assumption A.4 further justifies interchanging the derivative and integral on the left by Bartle (1966, corollary 5.9), so that

$$\int \mathsf{D}_j g(u \mid d, x)\ d\nu(u \mid x) = 0.$$

Substituting $\mathsf{D}_j g(u \mid d, x) = \mathsf{D}_j \log g(u \mid d, x)\ g(u \mid d, x)$ delivers the desired result. *(ii)* Given A.1*(ii)* and the conditions on τ_k, Theorem 3.1*(i)* gives $\mu_k(d, x) = \int \tau_k(r(d, z, u))$ $dG(u \mid d, x)$. Substituting $dG(u \mid d, x) = g(u \mid d, x)\ d\nu(u \mid x)$ gives

$$\mu_k(d, x) = \int \tau_k(r(d, z, u))\ g(u \mid d, x)\ d\nu(u \mid x).$$

Assumption A.6 permits application of Bartle (1966, corollary 5.9), ensuring differentiability of μ_k and the validity of an interchange of derivative and integral. Thus,

$$\begin{aligned} \mathsf{D}_j \mu_k(d, x) &= \int \mathsf{D}_j((\tau_k \circ r)g)(d, x, u)\ d\nu(u \mid x) \\ &= \int \mathsf{D}_j(\tau_k \circ r)(d, z, u)\ g(u \mid d, x)\ d\nu(u \mid x) \\ &\quad + \int \tau_k(r(d, z, u))\ \mathsf{D}_j g(u \mid d, x)\ d\nu(u \mid x), \end{aligned}$$

where A.1*(ii)* ensures the absence of terms involving $(\partial x / \partial d_j)$ in the second equality. Substituting $\mathsf{D}_j g(u \mid d, x) = \mathsf{D}_j \log g(u \mid d, x)\ g(u \mid d, x)$ delivers the result. *(iii)* The moment conditions ensure $|E(\mathsf{D}_j \tau_k(r(D, Z, U))\ s(D, X, U))| < \infty$ by Cauchy-Schwarz, ensuring the existence of $\delta_{1,k,j}(d, x)$ and thus of

$$\xi_{k,j}(d \mid x) = \int \mathsf{D}_j \tau_k(r(d, z, u))\ g(u \mid d, x)\ dv(u \mid x) - \delta_{1,k,j}(d, x).$$

The result now follows from *(ii)*. *(iv)* The result follows immediately from the Cauchy-Schwarz inequality, applied to *(iii)* and using *(i)*. ■

Proof of Theorem 4.3 (*i*) We write

$$\begin{aligned}\psi_\tau(d,x,m) \quad &:= \int \tau(r(d,z,u),m)\ dG(u \mid x) \\ &= \int \tau(r(d,z,u),m)[dG(u \mid x)\ /\ dG(u \mid d,x)]\ dG(u \mid d,x).\end{aligned}$$

The imposed second moment conditions ensure the existence and finiteness of this integral by Cauchy-Schwarz. (*ii*) The existence, uniqueness, and differentiability of ρ_0 follow immediately under the given conditions from the implicit function theorem (e.g., Chiang, 1984, pp. 210-211). (*iii*) Given the assumed differentiability of ψ_τ and Assumption A.7(*ii*), we have

$$\mathsf{D}_j\psi_\tau(d,x,\rho_0(d,x)) = \int \mathsf{D}_j\tau(r(d,z,u),\rho_0(d,x,))\ dG(u \mid x) = 0,$$

where the interchange of integral and derivative is justified by Bartle (1966, corollary 5.9), and the equality holds because $\psi_\tau(d,x,\rho_0(d,x)) = 0$ for all (d,x) in supp (D,X). Using the assumed differentiability of τ and r, the differentiability of ρ_0 ensured by (*ii*), and the chain rule gives

$$\int [\nabla_r\tau_\rho\ \mathsf{D}_j r + \nabla'_m\tau_\rho\ \mathsf{D}_j\rho_0]\ dG(u \mid x) = 0,$$

where we exploit the notation introduced preceding Theorem 4.3 in the text. Solving for $\mathsf{D}_j\rho_0$ given the assumed existence of $Q_\rho^{-1} = -[\int \nabla'_m\tau_\rho\ dG(u \mid x)]^{-1}$ yields

$$\mathsf{D}_j\rho_0 = Q_\rho^{-1}\int \nabla_r\tau_\rho\ \mathsf{D}_j r\ dG(u \mid x) = Q_\rho^{-1}\int \nabla_r\tau_\rho\ \mathsf{D}_j r\ g\ d\nu,$$

where the second equality holds given A.4(*ii*). (*iv*) A similar argument invoking A.4(*i*) and A.7(*i*) instead of A.4(*ii*) and A.7(*ii*) gives

$$\mathsf{D}_j\mu_0 = Q_\mu^{-1}\int [\nabla_r\tau_\mu\ \mathsf{D}_j r + \tau_\mu\ (\mathsf{D}_j\ \log g_d)]\ g_d\ d\nu,$$

given the assumed existence of $Q_\mu^{-1} = -[\int \nabla'_m\tau_\mu\ g_d\ d\nu]^{-1}$. It follows that

$$\mathsf{D}_j\mu_0 - \mathsf{D}_j\rho_0 = \int [Q_\mu^{-1}\nabla_r\tau_\mu - Q_\rho^{-1}\nabla_r\tau_\rho\ g/g_d]\ \mathsf{D}_j r\ g_d\ d\nu + Q_\mu^{-1}\int \tau_\mu(\mathsf{D}_j\ \log g_d)\ g_d\ d\nu.$$

The expression for $\delta_{0,j}$ holds by adding and subtracting terms appropriately. (*v*)(*a*)

$\delta_{1j}'\delta_{1j} = [\int \triangledown_r' \tau_\rho(\mathsf{D}_j r) s\ g_d\ dv]\ Q_\rho^{-1\prime} Q_\rho^{-1}\ [\int \triangledown_r \tau_\rho(\mathsf{D}_j r) s\ g_d\ dv] \leq \bar{\lambda}_\rho [\int \triangledown_r \tau_\rho(\mathsf{D}_j r) s\ g_d\ dv]'$ $[\int \nabla_r \tau_\rho(\mathsf{D}_j r)\ s\ g_d\ dv]$ by the Rayleigh inequality. The result then follows by Cauchy-Schwarz. (b) Analogous to (a). (c) Analogous to (a). ■

Acknowledgments: *The authors are indebted to Patrick Bajari, Clive W.J. Granger, Jinyong Hahn, Lars Hansen, James Heckman, Kei Hirano, Stefan Hoderlein, Kevin Hoover, Meng Huang, Charles Manski, Massimiliano Marinucci, Robert Marshall, Rosa Matzkin, Judea Pearl, Andres Santos, Chris Taber, and Rosalind Wu for helpful discussions and suggestions. Any errors remain the authors' responsibility.*

References

Abadie, A., J. Angrist, and G. Imbens (2002), "Instrumental Variables Estimates of the Effects of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91-117.

Abadie, A. and G. Imbens (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," NBER Technical Working Paper No. 283.

Ai, C. and Chen, X. (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795-1843.

Altonji, J. and R. Matzkin (2005), "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73, 1053-1102.

Angrist, J. and G. Imbens (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-476.

Angrist, J., G. Imbens, and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," (with Discussion), *Journal of the American Statistical Association*, 91, 444-455.

Barnow, B., G. Cain, and A. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," in E. Stromsdorfer and G. Farkas (eds.), *Evaluation Studies*, vol 5. San Francisco: Sage, pp. 43-59.

Bartle, R. (1966), *Elements of Integration.* New York: Wiley.

Berge, C. (1963), *Espaces Topologiques.* Paris: Dunod (translation by E.M. Patterson, *Topological Spaces.* Edinburgh: Oliver and Boyd).

Billingsley, P. (1979). *Probability Theory and Measure.* New York: Wiley.

Blundell, R. and J. Powell (2003), “Endogeneity in Nonparametric and Semiparametric Regression Models,” in M. Dewatripoint, L. Hansen, and S. Turnovsky (eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol II. New York: Cambridge University Press, pp. 312-357.

Brown, B. (1983), "The Identification Problem in Systems Nonlinear in the Variables," *Econometrica*, 51, 175-196.

Brown, D. and R. Matzkin (1998), "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," CFDP #1175, Yale University.

Chalak, K. and H. White (2007a), "An Extended Class of Instrumental Variables for the Estimation of Causal Effects," UCSD Department of Economics Discussion Paper.

Chalak, K. and H. White (2007b), “Independence and Conditional Independence in Causal Systems,” UCSD Department of Economics Discussion Paper.

Chalak, K. and H. White (2007c), “Identification with Conditioning Instruments in Causal Systems,” UCSD Department of Economics Discussion Paper.

Chen, X. (2005), "Large Sample Sieve Estimation of Semi-Nonparametric Models," New York University C.V. Starr Center Working Paper.

Chernozhukov, V. and C. Hansen (2005), "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245-261.

Chernozhukov, V., G. Imbens, and W. Newey (2007), "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4-14.

Chesher, A., (2003), "Identification in Nonseparable Models," *Econometrica*, 71, 1405-1441.

Chesher, A., (2005), "Nonparametric Identification under Discrete Variation," *Econometrica*, 73, 1525-1550.

Chiang, A. (1984). *Fundamental Methods of Mathematical Economics.* New York: McGraw-Hill.

Darolles, S., J. Florens, and E. Renault (2003), "Nonparametric Instrumental Regression," University of Tolouse GREMAQ Working Paper.

Das, M. (2005), "Instrumental Variables Estimators of Nonparametric Models with Discrete Endogenous Regressors," *Journal of Econometrics*, 124, 205-395.

Dawid, A.P. (1979), “Conditional Independence in Statistical Theory,” *Journal of the Royal Statistical Society*, Series B, 41, 1-31.

Dudley, R.M. (2002). *Real Analysis and Probability.* New York: Cambridge University Press.

Elliott, G. and R. Lieli (2005), "Predicting Binary Outcomes," UCSD Department of Economics Discussion Paper.

Engle, R., D. Hendry, and J.-F. Richard (1983), “Exogeneity,” *Econometrica*, 51, 277-304.

Firpo, S. (2007), "Efficient Semiparametric Estimation of Quantile Treatment Effects", *Econometrica*, 75, 259-276.

Firpo, S., N. Fortin, and T. Lemieux (2005), “Decomposing Wage Distributions: Estimation and Inference,” UBC Department of Economics Working Paper.

Fisher, F. (1966). *The Identification Problem in Econometrics.* New York: McGraw-Hill.

Fisher, F. (1970), “A Correspondence Principle for Simultaneous Equations Models,” *Econometrica*, 38, 73-92.

Goldberger, A. (1972), “Structural Equation Methods in the Social Sciences,” *Econometrica*, 40, 979-1001.

Grenander, U. (1981). *Abstract Inference.* New York: Wiley.

Hahn J. (1998), “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effect,” *Econometrica*, 66, 315-331.

Hahn, J. and G. Ridder (2007), "Conditional Moment Restrictions and Triangular Simultaneous Equations," IEPR Working Paper No. 07.3.

Hall, P. and J. Horowitz (2005), "Nonparametric methods for inference in the presence of instrumental variables," *Annals of Statistics*, 33, 2904-2929.

Hansen, L.P. (1982), “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50, 1029-1054.

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441-462.

Heckman, J. and R. Robb (1985), "Alternative Methods for Evaluating the Impact of Interventions," in J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labor*

Market Data. Cambridge: Cambridge University Press, pp. 146-245.

Heckman J. and E. Vytlacil (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences* 96, 4730-4734.

Heckman, J. and E. Vytlacil (2001), "“Local Instrumental Variables,” in C. Hsiao, K. Morimune, and J. Powell (eds.) in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*. Cambridge: Cambridge University Press, pp. 1-46.

Heckman, J. and E. Vytlacil (2005), “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73, 669-738.

Heckman, J. and E. Vytlacil (2007), "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin," University of Chicago, Department of Economics Discussion Paper.

Heckman, J., H. Ichimura, and P. Todd (1998), “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, 65, 261-294.

Heckman, J., J. Smith, and N. Clements (1997), “Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts,” *Review of Economic Studies*, 64, 487-535.

Heckman, J., S. Urzua, and E. Vytlacil (2006), "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88, 389-432.

Hirano, K. and G. Imbens (2001), "Estimation of Causal Effects using Propensity Score Weighting: An Application to Right Heart Catheterization," *Health Services and Outcomes Research*, 2, 259-278.

Hirano, K. and G. Imbens (2004), “The Propensity Score with Continuous Treatments,” in A. Gelman and X.-L. Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. New York: Wiley, pp. 73-84.

Hirano, K., G. Imbens, and G. Ridder (2003), “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, 71, 1161-1189.

Hoderlein, S. (2005), "Nonparametric Demand Systems, Instrumental Variables and a Heterogeneous Population," Mannheim University, Department of Economics Working Paper.

Hoderlein, S. (2007) "How Many Consumers are Rational?" Mannheim University, Department of Economics Working Paper.

Hoderlein, S. and E. Mammen (2007), "Identification of Marginal Effects in Nonseparable Models without Monotonicity," *Econometrica*, 75, 1513-1518.

Imbens, G. (2004), "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4-29.

Imbens, G. and W. Newey (2003), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," MIT Department of Economics Working Paper.

Lehmann, E. (1974). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

Li, Q., X. Lu, and A. Ullah (2003), "Multivariate local polynomial regression for estimating average derivatives," *Journal of Nonparametric Statistics*, 15, 607-624.

Lieli, R., and H. White (forthcoming), "The Construction of Empirical Credit Scoring Models Based on Maximization Principles," *Journal of Econometrics.*

Matzkin, R. (2003), "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339-1375.

Matzkin, R. (2004), "Unobservable Instruments," Northwestern University Department of Economics Working Paper.

Matzkin, R. (2005), "Identification of Nonparametric Simultaneous Equations," Northwestern University Department of Economics Working Paper.

Newey, W. and J. Powell (1988), "Instrumental Variables Estimation of Nonparametric Models," Manuscript, Department of Economics, Princeton University.

Newey, W. and J. Powell (2003), "Instrumental Variables Estimation of Nonparametric Models," *Econometrica*, 71, 1565-1578.

Newey, W., J. Powell, and F. Vella (1999), "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565-604.

Owen, A. (1988), "Empirical Likelihood Confidence Intervals for a Single Functional," *Biometrika*, 75, 237-249.

Owen, A. (2001). *Empirical Likelihood.* New York: CRC Press.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics.* Cambridge: Cambridge University Press.

Pearl, J. (1995), "Causal Diagrams for Experimental Research" (with Discussion), *Biometrika*, 82, 669-710.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press.

Powell, J., J. Stock, and T. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403-1430.

Ragusa, G. (2005), "Alternatives to GMM: Properties of Minimum Divergence Estimators," UCSD Department of Economics Discussion Paper.

Roehrig, C. (1988), "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433-447.

Rosenbaum, P. and D. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Santos, A. (2006), "Instrumental Variables Methods for Recovering Continuous Linear Functionals," Stanford University Department of Economics Working Paper.

Schennach, S. (2007), "Point Estimation with Exponentially Tilted Empirical Likelihood," *Annals of Statistics*, 35, 634-672.

Schennach, S., H. White, and K. Chalak (2007), "Estimating Average Marginal Effects in Nonseparable Structural Systems," UCSD Department of Economics Working paper.

Skouras, S. (2007), "Decisionmetrics: A Decision-Based Approach to Econometric Modeling," *Journal of Econometrics*, 137, 414-440.

Stoker, T. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1481.

Strotz, R. and H. Wold (1960), "Recursive vs. Nonrecursive Systems: An Attempt at Synthesis," *Econometrica*, 28, 417-427.

White, H. (2006), "Time Series Estimation of the Effects of Natural Experiments," *Journal of Econometrics,* 135, 527-566.

Wooldridge, J. (2002). *Econometric Analysis of Cross-Section and Panel Data.* Cambridge MA: MIT Press.