4-2017

Granger causality and structural causality in cross-section and panel data

Xun LU
Hong Kong University of Science and Technology

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

Halbert WHITE
University of California - San Diego

Citation

LU, Xun; SU, Liangjun; and WHITE, Halbert. Granger causality and structural causality in cross-section and panel data. (2017). *Econometric Theory*. 33, (2), 263-291.
Available at: https://ink.library.smu.edu.sg/soe_research/1961

Published in Econometric Theory, April 2017, Volume 33, Issue 2, Pages 263-291.
http://doi.org/10.1017/S0266466616000086

GRANGER CAUSALITY AND STRUCTURAL CAUSALITY IN CROSS-SECTION AND PANEL DATA

XUN LU
Hong Kong University of Science and Technology

LIANGJUN SU
Singapore Management University

HALBERT WHITE
University of California, San Diego

Granger noncausality in distribution is fundamentally a probabilistic conditional independence notion that can be applied not only to time series data but also to cross-section and panel data. In this paper, we provide a natural definition of structural causality in cross-section and panel data and forge a direct link between Granger ($G-$) causality and structural causality under a key conditional exogeneity assumption. To put it simply, when structural effects are well defined and identifiable, $G-$non-causality follows from structural noncausality, and with suitable conditions (e.g., separability or monotonicity), structural causality also implies $G-$causality. This justifies using tests of $G-$non-causality to test for structural noncausality under the key conditional exogeneity assumption for both cross-section and panel data. We pay special attention to heterogeneous populations, allowing both structural heterogeneity and distributional heterogeneity. Most of our results are obtained for the general case, without assuming linearity, monotonicity in observables or unobservables, or separability between observed and unobserved variables in the structural relations.

1. INTRODUCTION

Recently, White and Lu (2010, WL) have provided conditions establishing the equivalence of Granger ($G-$) causality and a natural notion of structural causality in structural vector autoregressions (VARs) and in time-series

This paper is dedicated to the memory of Hal White, who is profoundly missed. It builds on White's Econometric Theory Lecture at the 7th Symposium on Econometric Theory and Application (SETA, 2011) in Melbourne. White acknowledges Peter C. B. Phillips for his kind invitation. We gratefully thank the Editor Peter C. B. Phillips, the Co-editor Guido M. Kuersteiner, and three anonymous referees for their many constructive comments on the previous version of the paper. We are indebted to the participants of the SETA, who provided helpful suggestions and discussions. Su gratefully acknowledges the Singapore Ministry of Education for Academic Research Fund under grant number MOE2012-T2-2-021. All errors are the authors' sole responsibilities. Address correspondence to Xun Lu, Department of Economics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong; e-mail: xunlu@ust.hk.

natural experiments. The goal of this paper is to establish the analogous equivalence between $G-$causality and structural causality in cross-section and panel data under certain conditional exogeneity assumptions.

As $G-$causality is mostly examined in the time series context, it might be thought that it is strictly a time-series concept; if so, it would make no sense to talk about $G-$causality in cross-sections. In fact, however, $G-$causality is fundamentally a conditional independence notion, as pointed out by Florens and Mouchart (1982) and Florens and Fougère (1996). Holland (1986) states that "in my opinion, Granger's essential ideas involving causation do not require the time-series setting he adopted." As we show, $G-$causality has directly relevant and useful causal content not only for time-series cross-section panels, but also for pure cross-sections under certain conditional exogeneity assumptions.

In this paper, we focus on the aspects of the relation between $G-$causality and structural causality specific to cross-section or panel data. An important data feature here is unobserved heterogeneity. We pay special attention to two sources of heterogeneity that impact testing for structural causality, namely, structural heterogeneity and distributional heterogeneity. The structural heterogeneity refers to cross-group variation in unobservable constants (e.g., unknown nonrandom parameters) that enter the structural equation. Although unobserved heterogeneity has a long tradition in the literature on panel studies (see, Hsiao, 2014 and references therein), this type of heterogeneity with a group structure has recently received considerable attention (see, e.g., Sun, 2005, Lin and Ng, 2012, Deb and Trivedi, 2013, Lu and Su, 2014, Su, Shi, and Phillips, 2014, Bonhomme and Manresa, 2015, Sarafidis and Weber, 2015, and Bester and Hansen, 2016). The group structure has sound theoretical foundations from game theory or macroeconomic models where multiplicity of Nash equilibria is expected (see, e.g., Hahn and Moon, 2010). The distributional heterogeneity refers to the cross-group variation in certain conditional distributions, but seems to have received relatively less attention in the literature.[1] Browning and Carro (2007) provide some examples of such heterogeneity in micro-data. A similar question concerning distributional heterogeneity is also discussed in Hausman and Woutersen (2014) and Burda, Harding, and Hausman (2015) for duration models. The presence of either source of heterogeneity plays a central role in linking and testing $G-$causality and structural causality.

The main contributions of this paper can be clearly articulated. First, we introduce a heterogeneous population data generating process (DGP) for both cross-section and dynamic panel data and extend the concept of $G-$causality from the time series analysis to such settings. In particular, we focus on various versions of $G-$causality "in distribution" which are suitable for studying nonseparable and nonparametric structural equations. In the cross-section data, time is not explicitly involved, thus $G-$non-causality is a simple conditional independence relation. In panel data, time plays an explicit role and we pay special attention to the role of temporal precedence in defining $G-$causality.

Second, as in the time-series context, we give a natural definition of structural causality in cross-section and panel data. We distinguish the structural causality from various average causal effects. We show that given the conditional form of exogeneity, structural noncausality implies $G-$noncausality. If we further assume monotonicity or separability in the structural equations, structural causality implies $G-$causality. In the case where we do not assume monotonicity or separability, we strengthen structural causality to structural causality with positive probability ($w.p.p.$) and show that structural causality $w.p.p.$ implies $G-$causality. These results justify using tests of $G-$non-causality to test for structural noncausality. We emphasize that appropriately choosing covariates that ensure the conditional exogeneity assumption is the key to endowing $G-$non-causality with a structural interpretation. For example, we show that both leads and lags can be appropriate covariates in the panel data setting.

Third, we establish the linkage between population-group conditional exogeneity and its sample analogue in a heterogeneous population where the latter forms the basis for linking $G-$causality and structural causality. We show that without the conditional exogeneity at the sample level, the derivative of conditional expectation can be decomposed into three parts: a weighted average marginal effect, a bias term due to endogeneity, and a bias term due to heterogeneity. Thus, as emphasized in the literature (see, e.g., Angrist and Kuersteiner, 2004, 2011), without such assumptions as conditional exogeneity, it is impossible to give the $G-$non-causality test a causal interpretation. We show that conditional exogeneity ensures that the two bias terms vanish, in which case certain average subgroup causal effects with mixing weights are identified.

The plan of the paper is as follows. In Section 2, we provide a review on various concepts of causality. In Section 3, we first specify the cross-section heterogeneous population DGP and the sampling scheme. We define the cross-section structural causality and static $G-$causality and establish the equivalence of $G-$causality and structural causality under certain conditional exogeneity conditions. Testing for $G-$causality and structural causality is also discussed. In Section 4 we consider structural causality and $G-$causality in panel data. We focus on dynamic panel structures and derive some testable hypotheses. Section 5 concludes. All the mathematical proofs are gathered in the Appendix.

2. LITERATURE REVIEW

This paper builds on the vast literature on causality. For reviews on causality in econometrics, see Zellner (1979), Heckman (2000, 2008), Imbens (2004), Hoover (2008), Kuersteiner (2008), Angrist and Pischke (2009), Imbens and Wooldridge (2009), among others. Broadly speaking, we can divide the literature into two categories. The first includes $G-$causality and Sims causality. The second pertains to the causality defined on structural equations and that defined on potential outcomes.[2]

2.1. $G-$causality and Sims causality

$G-$causality and Sims causality were originally proposed to study time series data. Granger (1969) defines $G-$causality "in mean" based on conditional expectations, while Granger (1980) and Granger and Newbold (1986) generalize it to $G-$causality "in distribution" based on conditional distributions. In this paper, we focus on $G-$causality "in distribution", which we simply refer to as $G-$causality. To define it, we first introduce some notation. For any sequence of random vectors $\{Y_t,\ t = 0, 1, ...\}$, we let $Y^t \equiv (Y_0, ..., Y_t)$ denote its "$t-$history", and let $\sigma(Y^t)$ denote the sigma-field generated by Y^t. Let $\{D_t, Y_t, X_t\}$ be a sequence of random vectors, where $\{D_t\}$ is the cause of interest, $\{Y_t\}$ is the response of interest, and $\{X_t\}$ is some variates. Granger and Newbold (1986) say that D_{t-1} does not $G-$cause Y_{t+k} with respect to $\sigma(Y^{t-1}, X^{t-1})$ if for all $t = 1, 2, ...,$

$$F_{t+k}\left(\ \cdot\ \middle|\ D^{t-1}, Y^{t-1}, X^{t-1}\right) = F_{t+k}\left(\ \cdot\ \middle|\ Y^{t-1}, X^{t-1}\right), \quad k = 0, 1, ..., \tag{2.1}$$

where $F_{t+k}\left(\ \cdot\ |\ D^{t-1}, Y^{t-1}, X^{t-1}\right)$ denotes the conditional distribution function of Y_{t+k} given $\left(D^{t-1}, Y^{t-1}, X^{t-1}\right)$, and $F_{t+k}\left(\ \cdot\ |\ Y^{t-1}, X^{t-1}\right)$ denotes that of Y_{t+k} given $\left(Y^{t-1},\ X^{t-1}\right)$. As Florens and Mouchart (1982) and Florens and Fougère (1996) have noted, eq.(2.1) is equivalent to the conditional independence relation:

$$Y_{t+k} \perp D^{t-1}\ |Y^{t-1}, X^{t-1}, \tag{2.2}$$

where we use $\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}$ to denote that $\mathcal{X}$ and $\mathcal{Y}$ are independent given $\mathcal{Z}$. If

$$E\left(Y_{t+k}\middle|D^{t-1}, Y^{t-1}, X^{t-1}\right) = E\left(Y_{t+k}\middle|Y^{t-1}, X^{t-1}\right)\ a.s., \tag{2.3}$$

we say that D_{t-1} does not $G-$cause Y_{t+k} with respect to $\sigma\left(Y^{t-1}, X^{t-1}\right)$ "in mean". Eq.(2.3) is a conditional mean independence statement. In the literature, most of the discussion on $G-$causality focuses on $k = 0$ (see, e.g., Granger, 1980, 1988, and Kuersteiner, 2008). The definition of $G-$causality does not rely on any economic theory or structural assumptions. Granger (1969, p. 430) emphasizes that "The definition of causality used above is based entirely on the predictability of some series." Therefore, $G-$causality does not necessarily reveal any underlying causal relation, and it is entirely unfounded to draw any *structural* or policy conclusions from the $G-$causality tests.

"Sims noncausality" was originally defined in Sims (1972). Florens (2003) and Angrist and Kuersteiner (2004, 2011) give a generalized definition:

$$Y_t^{\infty} \perp D_{t-1}\ |D^{t-2}, Y^{t-1}, X^{t-1},\ \ t = 1, ...,$$

where $Y_t^{\infty} = (Y_t, Y_{t+1}, ...)$. Chamberlain (1982) and Florens and Mouchart (1982) show that under some mild regularity conditions, $G-$noncausality with $k = 0$ and Sims noncausality are equivalent when the covariates $\{X_t\}$ are absent. When the covariates $\{X_t\}$ are present, however, $G-$non-causality and Sims noncausality

are in general not equivalent. For an excellent review on the relationship between $G-$noncausality and Sims noncausality, see Kuersteiner (2008). Similar to $G-$noncausality, Sims noncausality is completely based on predictability and has no *structural* interpretation.

2.2. Structural causality and causality in the potential outcome framework

On the other hand, causality defined on structural equations or potential outcomes is mostly discussed in cross-section data. The structural equation approach can be traced back to the work of the Cowles Commission in the 1940s (see, e.g., Haavelmo, 1943, 1944, and Koopmans, 1950), though most of their work was based on linear equations. More recently, researchers have generalized linear equations to nonseparable and nonparametric equations (see, e.g., Chesher, 2003, 2005, Matzkin 2003, 2007, and Altonji and Matzkin, 2005). In 2015, *Econometric Theory* published a special issue in memory of Trygve Haavelmo (Volume 31, Issue 01 and 02) which contains many interesting discussions on recent developments of causality related to Haavelmo's structural models.

Consider a simple case where the response of interest is Y_i and the cause of interest is D_i. The causal relation between D_i and Y_i can be characterized by an unknown structural equation r such that

$$Y_i = r(D_i, U_i),$$

where U_i is other unobservable causes of Y_i. For example, when Y_i is the demand, D_i is the price and U_i represents certain demand shocks, r is the demand function that is derived from utility maximization. r is a general function and does not need to be linear, parametric, or separable between observable causes D_i and unobservable causes U_i. Here r has a *structural* or *causal* meaning and we define the causal effect of D_i on Y_i based on r. Let $\mathbb{D}$ and $\mathbb{U}$ be the support of D_i and U_i, respectively. If $r(d, u)$ is a constant function of d for all $d \in \mathbb{D}$ and all $u \in \mathbb{U}$, then we simply say that D_i does not structurally cause Y_i (see, e.g., Heckman, 2008, and White and Chalak, 2009). For a binary D_i, the effect of D_i on Y_i is $r(1, u) - r(0, u)$ when $U_i = u$. The effect can depend on unobservable u, thus unobservable heterogeneity is allowed. For a continuous D_i, the marginal effect of D_i on Y_i is $\partial r(d, u)/\partial d$ when $D_i = d$ and $U_i = u$. To identify the effects of D_i on Y_i, we often impose the assumption of *conditional exogeneity*:

$$D_i \perp U_i \mid X_i,$$

where X_i is some observable covariates. This includes the special case where D_i and U_i are independent, corresponding to D_i being randomized.

Halbert White has made substantial contribution on defining, identifying and estimating causal effects in structural equations. White and Chalak (2009) extend Judea Pearl's causal model to a settable system which incorporates features

of central interest to economists and econometricians: optimization, equilibrium, and learning. Roughly speaking, a settable system is "a mathematical framework describing an environment in which multiple agents interact under uncertainty" (White and Chalak, 2009, p. 1760). In the settable system, a variable of interest has two roles: "responses" and "settings". When the value of the variable is determined by the structural equation, these values are called "responses". In contrast, when the value is not determined by the structural equation, but is instead set to one of its admissible values, these values are called "settings". They show that on the settable system, causes and effects can be rigorously defined. Chalak and White (2012) provide definitions of direct, indirect and total causality on the settable system, in terms of functional dependence, and show how causal relations and conditional independence are connected. Chalak and White (2011) provide an exhaustive characterization of potentially identifying conditional exogeneity relationships in liner structural equations and introduce conditioning and conditional extended instrumental variables to identify causal effects. White and Chalak (2013) provide a detailed discussion on the identification and identification failure of causal effects in structural equations. White and Chalak (2010) discuss how to test the conditional exogeneity assumption.

The treatment effect literature adopts the potential outcome framework (see, e.g., Rubin, 1974, Rosenbaum and Rubin, 1983, and Holland, 1986). For a cause of interest D_i and response of interest Y_i, we define a collection of potential outcomes $\{Y_i(d), d \in \mathbb{D}\}$, where $\mathbb{D}$ is the support of D_i. Then we can define causal or treatment effect based on the potential outcomes. Specifically, for two values d and d^* in $\mathbb{D}$, we define $Y_i(d^*) - Y_i(d)$ as the causal effects of D_i on Y_i when D_i changes from d to d^*. If $Y_i(d) = Y_i(d^*)$ for all d and $d^* \in \mathbb{D}$, we say that D_i has no causal effects on Y_i. For example, when D_i is binary, i.e., $\mathbb{D} = \{0, 1\}$, $Y_i(0)$ and $Y_i(1)$ are the potential outcomes, corresponding $D_i = 0$ and $D_i = 1$, respectively. The treatment or causal effect is $Y_i(1) - Y_i(0)$. Two average effects often discussed in the literature are

the average treatment effect (ATE) : $E(Y_i(1) - Y_i(0))$,

and

the average treatment effect on the treated (ATT) : $E(Y_i(1) - Y_i(0) \mid D_i = 1)$.

Note that we only observe one outcome in data. For example, when D_i is binary, the observed outcome is simply $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. To identify the effects, we often impose the assumption of *unconfoundedness* or *selection on observables*:

$Y_i(d) \perp D_i \mid X_i$,

where X_i is observable covariates. Lechner (2001), Imbens (2004), Angrist and Pischke (2009), and Imbens and Wooldridge (2009) provide reviews on identification and estimation of various effects in the potential outcome framework.

White and Chalak (2013, p. 280) show that the structural equation framework is equivalent to the potential outcome framework. For example, the potential outcome $Y_i(d)$ is simply $r(d, U_i)$. The unconfoundedness assumption $Y_i(d) \perp D_i \mid X_i$ in the potential outcome framework is equivalent to the conditional exogeneity assumption $U_i \perp D_i \mid X_i$ in the structural equation framework.

2.3. Relationship between these two types of causality

We emphasize that $G-$causality and Sims causality are entirely based on predictability, while the causality defined on structural equations or potential outcomes is a real causal relation. In the literature, there are several papers that link these two types of causality. White and Lu (2011) provide a direct link between $G-$causality and structural causality in structural equations. They show that given the conditional exogeneity assumption, structural noncausality is essentially equivalent to $G-$noncausality. Angrist and Kuersteiner (2004, 2011) and Kuersteiner (2008) discuss the link between Sims noncausality and the noncausality defined in the potential outcome framework under the assumption of selection on observables. Angrist, Jordà, and Kuersteiner (2013) study the monetary policy effects in the potential outcome framework. Lechner (2011) links Granger/Sims noncausality to various average effects defined in the potential outcome framework. However, all the discussions so far have been made in the context of time series data.

3. STRUCTURAL CAUSALITY AND $G-$CAUSALITY IN CROSS-SECTION DATA

3.1. Structural causality

We first specify a population DGP for a single time period, omitting the time index. We write $\mathbb{N}^+ := \{1, 2, ...\}$ and $\mathbb{N} := \{0\} \cup \mathbb{N}^+$. We also write $\bar{\mathbb{N}}^+ := \mathbb{N}^+ \cup \{\infty\}$ and $\bar{\mathbb{N}} := \{0\} \cup \bar{\mathbb{N}}^+$.

Assumption A.1 (Cross-Section Heterogeneous Population DGP)**.** Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, let $N \in \mathbb{N}^+$, and for the members of a population $j \in \{1, ..., N\}$ let the random vector Y_j be structurally determined by a triangular system as

$$Y_j = r\left(D_j, Z_j, U_j; b_j\right), \tag{3.1}$$

where r is an unknown measurable $k_y \times 1$ function, $k_y \in \mathbb{N}^+$; D_j, Z_j, and U_j, $j = 1, 2, ...$, are vectors of nondegenerate random variables on $(\Omega, \mathcal{F}, P)$ having dimensions $k_d \in \mathbb{N}^+$, $k_z \in \mathbb{N}$, and $k_u \in \bar{\mathbb{N}}$, respectively; and b_j is a nonrandom real vector of dimension $k_b \in \bar{\mathbb{N}}$. Suppose also that W_j is a random vector on $(\Omega, \mathcal{F}, P)$ with dimension $k_w \in \mathbb{N}$. The triangular structure is such that Y_j does not structurally determine D_j, Z_j, and U_j, and neither Y_j nor D_j structurally determines W_j.

To keep causal concepts clear, it is important to distinguish between the population DGP, defined above, and the sample DGP, defined later. For now, we leave sampling aside. We may also refer to members of the population as "units" or "individuals" and the distribution of the population as a mixture distribution. Mixture distributions appear in many contexts in the literature and arise naturally where a statistical population contains two or more subpopulations (see, e.g., Lindsay, 1995, and Mclachlan and Basford, 1988). As will be clear in a moment, the sources of population heterogeneity are in general not observed.

We interpret Y_j as the response of interest. It is determined by variables D_j, Z_j, and U_j, and the constants b_j. The variables can be binary, categorical, or continuous. Immediately below, we formalize a natural notion of causality for this system. For now, it is heuristically appropriate to view D_j as observable causes of interest (e.g., a treatment), Z_j as other observable causes not of primary interest, and U_j as unobservable causes.[3]

The structural function r is unknown, and our goal will be to learn about it from a sample of the population. Nevertheless, when W_j has positive dimension, r embodies the *a priori* exclusion restriction that W_j does not determine Y_j. The typical sources of this restriction, as well as the identities of D_j, Z_j, and U_j, their status as observable or not, and the priority or precedence relations embodied in the assumed triangularity, are economic theory and specific domain knowledge.

The assumed triangularity rules out explicit simultaneity for succinctness and clarity. For the causality in simultaneous structural equations, there are different views in the literature. White and Chalak (2009, 2013) follow Strotz and Wold (1960) and argue that simultaneous equations are not causal structural relations, but are instead "mutual consistency conditions holding between distinct sets of structural equations – for example, between one set of structural equations governing partial equilibrium and another governing full equilibrium". However, other researchers argue that simultaneous structural equations can be given a causal interpretation. For example, Angrist, Graddy, and Imbens (2000) provide a potential outcome interpretation for general nonlinear simultaneous equation models. They also show that the standard linear instrumental variable estimator identifies a weighted average of the causal effects.

By definition, the constants b_j are fixed for a given individual. Nevertheless, they may vary across individuals; we call this variation *structural heterogeneity*. If the b_j's are identical, we write them as b_0. We assume that b_j's are unknown. Note that the presence of b_j facilitates writing r without a j subscript, as differences in the structural relations across population members can be accommodated by variations in the possibly infinite-dimensional b_j.

To give a definition of causality for the structural system in Assumption A.1, let $\mathbb{D}_j := \text{supp}(D_j)$ denote the support of D_j, i.e., the smallest closed set containing D_j with probability 1.

DEFINITION 3.1 (Cross-Section Structural Causality). Let j be given. If the function $r(\cdot, z, u; b_j) : \mathbb{D}_j \to \mathbb{R}^{k_y}$ is constant on $\mathbb{D}_j$ for all admissible z and u,

then D_j ***does not structurally cause*** Y_j, and we write $D_j \not\Rightarrow_{\mathcal{S}} Y_j$. Otherwise, D_j ***structurally causes*** Y_j, and we write $D_j \Rightarrow_{\mathcal{S}} Y_j$.

Here we implicitly assume that there is variation in potential cause D_j (i.e., D_j is nondegenerated) and under the set of counterfactual policies $\mathbb{D}_j$, the structural function r is invariant. Therefore, the invariant function r fully characterizes the causal/structural relationship between Y_j and (D_j, Z_j, U_j).

Structural causality is structural functional dependence in a specific context. Similar definitions can be given for Z_j and U_j, but we leave these implicit, as these are not the main causes of interest. Causality for components of D_j is defined in an obvious way. If we define $Y_j(d) := r\left(d, Z_j, U_j; b_j\right)$ as the potential outcomes (see, e.g., Rubin, 1974, Rosenbaum and Rubin, 1983, and Holland, 1986), then the equivalent definition of structural noncausality is that $Y_j(d) = Y_j(d^*)$ *a.s.* for all $(d, d^*) \in \mathbb{D}_j$.

Structural causality can be easily understood in the familiar linear structure for scalar Y_j,

$$Y_j = b_{j,0} + D_j' b_{j,1} + Z_j' b_{j,2} + U_j,$$

where $b_j := (b_{j,0}, b_{j,1}', b_{j,2}')'$. If $b_{j,1} = 0$, then D_j does not structurally cause Y_j. Otherwise, it does. This is so natural and intuitive that one might wonder why causality is such a thorny topic. One main reason for confusion arises from the relation between causality and simultaneity as discussed above; the triangular systems considered here obviate this issue. Another main reason for confusion is the failure to distinguish carefully between structural equations of the sort written above and regression equations, which may look similar but need not have structural content. The equation above is entirely structural. We will encounter regressions (conditional expectations) only after suitable structural foundations are in place.

Observe that heterogeneity in b_j's permits D_j to cause Y_j for some j and not for others.

The nondegeneracy of D_j ensures that $\mathbb{D}_j$ contains at least two points, so D_j is a variable rather than a constant. Variation in potential causes is fundamental for defining causality (c.f. Holland, 1986); this is what makes possible analogous definitions of causality for Z_j and U_j. Significantly, however, because b_j's are fixed constants, they cannot be causes of Y_j. Instead, b_j's can be effects or can determine effects. This follows from the following formal definition:

DEFINITION 3.2 (Intervention and Effect). Let j be given, and let d and d^* be distinct admissible values for D_j. Then $d \to d^*$ is an ***intervention*** to D_j. Let (z, u) be admissible values for (Z_j, U_j). The difference $y_j^* - y_j = r\left(d^*, z, u; b_j\right) - r\left(d, z, u; b_j\right)$ is the ***effect on*** Y_j ***of the intervention*** $d \to d^*$ ***to*** D_j ***at*** (z, u).

This is also referred to as the "causal effect" in the treatment effect literature (see, e.g., Rubin, 1974). For the linear case with scalar d, the effect of a one-unit

intervention $d \to d+1$ is

$$y_j^* - y_j = \left(b_{j,0} + (d+1)b_{j,1} + z'b_{j,2} + u\right) - \left(b_{j,0} + d\,b_{j,1} + z'b_{j,2} + u\right) = b_{j,1}.$$

For unit j, this effect is fixed; i.e., it is the same constant for all (z, u). Because b_j can differ across units, this permits effect heterogeneity.

Alternatively, if $Y_j = b_{j,0} + Z_j D_j' b_{j,1} + U_j$, with scalar Z_j, then the effect of $d \to d^*$ is

$$y_j^* - y_j = \left(b_{j,0} + zd^{*\prime}b_{j,1} + u\right) - \left(b_{j,0} + zd'b_{j,1} + u\right) = z\left(d^* - d\right)' b_{j,1}.$$

Here, the effect of D_j on Y_j depends on z but not u. In this case, $b_{j,1}$ determines the effect of D_j on Y_j, together with Z_j. When an effect depends on a variable (i.e., an element of Z_j or U_j), it is standard (especially in the epidemiological literature) to call that variable an *effect modifier*.

For simplicity, we call $y_j^* - y_j$ the effect of $d \to d^*$. Chalak and White (2012) discuss indirect, direct, and total effects in structural equations. In the potential outcome framework, Rubin (2004) discusses "direct" and "indirect" casual effect using principal stratification.

We distinguish the structural causal effect from various average causal effects, such as $E[r(d^*, Z_j, U_j;\ b_j) - r(d, Z_j, U_j;\ b_j)]$. It is clear that structural non-causality implies zero average causal effects, while the converse is not necessarily true.

To describe identified effects in samples from heterogeneous populations, we define population *groups* $\mathcal{J}_{\boldsymbol{g}}$, $\boldsymbol{g} = 1, ..., \Gamma$, as collections of population units having identical b_j and identical distributions of $(D_j, U_j) \mid X_j$, where $X_j := (Z_j', W_j')'$. We define $\mathcal{G} := \{1, ..., \Gamma\}$. As the population is finite, so is the number of groups: $\Gamma \leq N$. We define $N_{\boldsymbol{g}} := \#\{j \in \mathcal{J}_{\boldsymbol{g}}\}$, where $\#\{\cdot\}$ is the cardinality of the indicated set, and let $p_{\boldsymbol{g}} := N_{\boldsymbol{g}}/N$ be the proportion of population units belonging to group $\mathcal{J}_{\boldsymbol{g}}$. The b_j's by themselves need not define groups, as the distributions of $(D_j, U_j) \mid X_j$ may differ for units with identical b_j. We call cross-group variation in the distributions of $(D_j, U_j) \mid X_j$ *distributional heterogeneity*, to distinguish this from *structural heterogeneity* that refers to cross-group variation in b_j.

Note that because groups are defined by unobservable constants, b_j, and distributions involving the unobservable U_j, in general, we do not know for sure whether two units belong to the same group. Interestingly, in the related literature, Bester and Hansen (2016) consider estimating grouped effects in panel data models when each individual's group identity is known, and Su, Shi, and Phillips (2014) consider identifying latent group structures in panel data models via a variant of Lasso. Nevertheless, both groups of researchers have only focused on the structural heterogeneity in the panel framework.

Given A.1, it follows that $(Y_j, D_j) \mid X_j$ is identically distributed for all units j in group $\mathcal{J}_{\boldsymbol{g}}$. As a convenient shorthand, for $j \in \mathcal{J}_{\boldsymbol{g}}$, we write $(Y_j, D_j) \mid X_j \sim (Y_{\boldsymbol{g}}, D_{\boldsymbol{g}}) \mid X_{\boldsymbol{g}}$, where $A \sim B$ denotes that A is distributed as B.

Typically, we do not observe an entire population. Instead we observe observations sampled from the population in some way. For simplicity and concreteness, we consider simple random sampling. At the sample level, we write[4]

$$Y = r(D, Z, U; B),$$

where B is the randomly sampled b_j. Note that the randomness of B arises *solely* from the sampling process. We write $X := (Z', W')'$ and distinguish X from the causes D by calling X a vector of *covariates*.

For any random vector X, we let $f(x)$ and $F(x)$ denote the joint probability density function (PDF) and cumulative distribution function (CDF), respectively. For any two random vectors X and Y, we use $f(y|x)$ and $F(y|x)$ to denote the conditional PDF and CDF of Y given $X = x$, respectively. We also use subscript j and $\boldsymbol{g}$ to denote individual j and group $\boldsymbol{g}$, respectively. For example, $F_{\boldsymbol{g}}(u, d, x)$ and $f_{\boldsymbol{g}}(u|d, x)$ denotes the CDF of $(U_{\boldsymbol{g}}, D_{\boldsymbol{g}}, X_{\boldsymbol{g}})$ and the conditional PDF of $U_{\boldsymbol{g}}$ given $D_{\boldsymbol{g}}, X_{\boldsymbol{g}}$ for group $\boldsymbol{g}$, respectively. Note that $F_{\boldsymbol{g}}(u, d, x) := N_{\boldsymbol{g}}^{-1} \sum_{j \in \mathcal{J}_{\boldsymbol{g}}} F_j(u, d, x)$, which defines the *mixture* CDF of $(U_{\boldsymbol{g}}, D_{\boldsymbol{g}}, X_{\boldsymbol{g}})$.

3.2. $G-$causality: a first encounter

We now define $G-$causality for cross-section data based on Granger's philosophy of nonpredictability. Holland (1986, p. 957) considers a special case where D is randomized and applies $G-$causality to cross-section data.[5] As in Holland, we define $G-$non-causality as a conditional independent statement. Following Dawid (1979), we write $\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}$ to denote that $\mathcal{X}$ and $\mathcal{Y}$ are independent given $\mathcal{Z}$ and $\mathcal{X} \not\perp \mathcal{Y} \mid \mathcal{Z}$ if $\mathcal{X}$ and $\mathcal{Y}$ are not independent given $\mathcal{Z}$. Translating Granger and Newbold's (1986, p. 221) definition to the cross-section context gives

DEFINITION 3.3 ($G-$Causality). Let Y, Q, and S be random *vectors*. *If* $Q \perp Y \mid S$, then Q ***does not*** $G-$***cause*** Y ***with respect to (w.r.t.)*** S. Otherwise, Q $G-$***causes*** Y ***w.r.t.*** S.

Here Y, Q, and S can be any random vectors, and this notion has no structural content. When S is a constant, we have the simplest form of $G-$non-causality: independence. Correlation is an example of $G-$causality.

To give $G-$causality structural meaning, we impose A.1, assume that D obeys a suitable exogeneity condition, and take $Q = D$ and $S = X$. The exogeneity assumed for D ensures identification for various measures of its effects and can often be structurally justified. It also turns out to be ideally suited to endowing $G-$causality with structural meaning.

To see how this works in a simple setting, consider the homogeneous case with D being a binary scalar, in which average treatment effects (ATE) and average treatment effects on the treated (ATT) are often discussed (see, e.g., Rubin, 1974,

Hahn, 1998, Hirano, Imbens and Ridder, 2003 and Angrist and Pischke, 2009). Define the covariate-conditioned effect of treatment on the treated as

$$ATT_X := E\left(Y(1) - Y(0) \mid D = 1, X\right),$$

where $Y(1)$ is the potential response to treatment and $Y(0)$ is the potential response in the absence of treatment. Using A.1, we write $Y(1) := r(1, Z, U; b_0)$ and $Y(0) := r(0, Z, U; b_0)$. ATT_X thus has a clear structural interpretation. From this, we can construct ATT,

$$ATT := E\left(Y(1) - Y(0) \mid D = 1\right) = E(ATT_X \mid D = 1).$$

Note that here we define ATT and ATT_X based on our structural equation r. Nevertheless, they can be defined without a structural model, i.e., based on the two potential outcomes $Y(0)$ and $Y(1)$ (see, e.g., Rubin, 1974 and Angrist and Pischke, 2009).

To identify ATT_X, it suffices that $D \perp Y(0) \mid X$. Given A.1, it suffices for this that $D \perp U \mid X$, as is readily verified. This *conditional exogeneity* is a common identifying assumption in cross-sections.[6] Classical strict exogeneity ($(D, Z) \perp U$) is sufficient but not necessary for this. Using $D \perp U \mid X$ in the second line below, we have

$$\begin{aligned} ATT_X &= E\left[r(1, Z, U; b_0) \mid D = 1, X\right] - E\left[r(0, Z, U; b_0) \mid D = 1, X\right] \\ &= E\left[r(1, Z, U; b_0) \mid D = 1, X\right] - E\left[r(0, Z, U; b_0) \mid D = 0, X\right] \\ &= E(Y \mid D = 1, X) - E(Y \mid D = 0, X) =: \mu(1, X) - \mu(0, X). \end{aligned}$$

Thus, ATT_X can be expressed in terms of the distribution of observables (here $\mu(1, X)$ and $\mu(0, X)$), so this effect is *identified* (e.g., Hurwicz, 1950). Similarly, ATT is identified. The identification result based on conditional exogeneity has been extensively discussed in the treatment effect literature (see, e.g., Rubin, 1974, Holland, 1986 and Angrist and Pischke, 2009).

To see the structural meaning for $G-$causality, observe that $E(Y \mid D = 1, X) - E(Y \mid D = 0, X) = 0$ *a.s.* is another way to write $E(Y \mid D, X) = E(Y \mid X)$ *a.s.* Thus, in this context, $G-$non-causality in mean of D for Y w.r.t. X is *equivalent* to $ATT_X = 0$ *a.s.*

Note that, even for binary treatments, $G-$causality in mean does not tell the full story, as $G-$causality in distribution can hold even when ATT_X or ATT vanishes. For example, suppose

$$Y = UD,$$

with $D \perp U$, $E(U) = 0$, $E(U^2) < \infty$, and $E(D^2) < \infty$. Then D structurally causes Y; its effect on Y is U, a random effect. These effects may be positive or negative; however, they average out, as $ATT = 0$.[7] Here, and as we will show generally, $G-$non-causality essentially has the interpretation that Y is not

structurally caused by D under the key conditional exogeneity assumption. In this example, we have $G-$causality ($D \not\perp Y$), as $E(Y^2 \mid D) = D^2 E(U^2)$. Testing $G-$non-causality will detect the structural causality here, in contrast to testing $ATT = 0$ or, equivalently, $G-$non-causality in mean of D for Y.

When D is nonbinary, A.1 together with suitable exogeneity conditions, similarly suffices to identify certain effects that give structural meaning to $G-$causality. This makes it generally possible to test for structural causality by testing for $G-$causality.

With heterogeneity, matters become somewhat more complicated. In particular, in the binary case, $\mu(1, X) - \mu(0, X)$ is no longer ATT_X. Instead, with suitable exogeneity (Assumption A.2 below), we have that for each x, $\mu(1, x) - \mu(0, x)$ recovers a *blended* effect

$$\mu(1, x) - \mu(0, x) = \sum_{\boldsymbol{g}=1}^{\Gamma} \zeta_{\boldsymbol{g}|x} \, ATT_{\boldsymbol{g}|x},$$

where the $\zeta_{\boldsymbol{g}|x}$'s are nonnegative weights adding to one, and for all $j \in \mathcal{J}_{\boldsymbol{g}}$, $ATT_{\boldsymbol{g}|x} := E[Y_j(1) - Y_j(0) \mid D_j = 1, X_j = x]$, is the group-specific covariate-conditioned average effect of treatment on the treated. Here, $Y_j(1)$ is the potential response to treatment and $Y_j(0)$ is the potential response in the absence of treatment.

3.3. Exogeneity and effect identification with heterogeneity

With heterogeneity, an identifying exogeneity condition relevant for analyzing $G-$causality is

Assumption A.2 (Heterogeneous Cross-Section Exogeneity)**.** For all $\boldsymbol{g} \in \mathcal{G}$, (i) $D_{\boldsymbol{g}} \mid X_{\boldsymbol{g}} \sim D \mid X$; and (ii) $D_{\boldsymbol{g}} \perp U_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$.

Observe that this imposes structures for all groups, namely, all $\boldsymbol{g} \in \mathcal{G}$. Although this condition is not the weakest possible (see Chalak and White (2012) and the discussion below), A.2 generally suffices to identify effects and link $G-$causality and structural causality. A.2(i) restricts the allowed distributional heterogeneity: for all $\boldsymbol{g} \in \mathcal{G}$, the conditional distributions of $D_{\boldsymbol{g}}$ given $X_{\boldsymbol{g}}$ are assumed identical. A.2(ii) imposes conditional exogeneity for all groups. Below, we discuss this further.

THEOREM 3.1. *Suppose A.1 - A.2*(i) *hold. If A.2*(ii) *holds, then* $D \perp U \mid X$.

It is also of interest to know whether the converse holds. A strict converse does not hold, as certain fortuitous cancellations in population-group conditional distributions can yield sample conditional exogeneity without population-group conditional exogeneity. Nevertheless, the converse does hold under a mild regularity condition ruling out exceptional cases. We introduce the following definition.

DEFINITION 3.4. $\{U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ is ***regular*** if

$$\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d, x)\, f_{\boldsymbol{g}}(x) = \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid x) f_{\boldsymbol{g}}(x) \text{ for all } (u, d, x) \tag{3.2}$$

implies $U_{\boldsymbol{g}} \perp D_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$ for all $\boldsymbol{g} \in \mathcal{G}$.

When $\#\mathcal{G} = 1$, $\{U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ is necessarily regular. To understand what regularity rules out, consider the next simplest case, with $\#\mathcal{G} = 2$ and $X_{\boldsymbol{g}}$ absent. Then $\{U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ fails to be regular if and only if $U_{\mathbf{2}} \not\perp D_{\mathbf{2}}$ (say) and

$$f_{\mathbf{1}}(u \mid d) = f_{\mathbf{1}}(u) - \frac{1 - p_{\mathbf{1}}}{p_{\mathbf{1}}} [f_{\mathbf{2}}(u \mid d) - f_{\mathbf{2}}(u)] \text{ for all } (u, d),$$

where the two groups have been re-indexed for convenience. Suppose that $p_{\mathbf{1}}$, $f_{\mathbf{1}}(u)$, $f_{\mathbf{2}}(u)$, and $f_{\mathbf{2}}(u \mid d)$ are arbitrary. Then it is easily arranged that $f_{\mathbf{1}}(u \mid d)$ can be negative for (u, d) in a set of positive probability, so $f_{\mathbf{1}}(u \mid d)$ does not define a conditional density, and $\{U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ is regular after all. If this $f_{\mathbf{1}}(u \mid d)$ is nevertheless a conditional density, it is clearly highly special, as it ensures that eq.(3.2) holds for the given $p_{\mathbf{1}}$ and the functions defined by $f_{\mathbf{1}}(u)$, $f_{\mathbf{2}}(u)$, and $f_{\mathbf{2}}(u \mid d)$ (all of which are typically unknown), but not necessarily otherwise.

In the general case, $\{U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ fails to be regular if and only if $U_{\boldsymbol{g}} \not\perp D_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$ for some $\boldsymbol{g} \in \mathcal{G}$ and

$$f_{\mathbf{1}}(u \mid d, x) = f_{\mathbf{1}}(u \mid x) - (p_{\mathbf{1}} f_{\mathbf{1}}(x))^{-1} \sum_{\boldsymbol{g}=\mathbf{2}}^{\Gamma} p_{\boldsymbol{g}} [f_{\boldsymbol{g}}(u \mid d, x) - f_{\boldsymbol{g}}(u \mid x)] f_{\boldsymbol{g}}(x)$$
$$\times \text{ for all } (u, d, x),$$

where we again re-index the groups. As before, this is clearly a very special population configuration; ruling out such cases by imposing regularity is a weak restriction.

The converse result is

THEOREM 3.2. *Suppose A.1 - A.2*(i) *hold, and suppose* $\{U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ *is regular. If* $D \perp U \mid X$, *then A.2*(ii) *holds.*

Theorems 3.1-3.2 also hold with $\{Y_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ regular, $D \perp U \mid X$ replaced by $D \perp Y \mid X$, and A.2(ii) replaced by $D_{\boldsymbol{g}} \perp Y_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$, $\boldsymbol{g} \in \mathcal{G}$. This result plays a central role in linking structural causality and $G-$causality.

3.4. Linking structural causality and $G-$causality

As we have seen, A.1 imposes structures where causal effects are well defined; A.2 permits recovering versions of these. As we now prove, this gives structural meaning to $G-$causality generally. Recall that the b_j's are identical in group $\mathcal{J}_{\boldsymbol{g}}$; thus, the same structural causality relations hold for all j in a given group. We write $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}$ when $D_j \not\Rightarrow_{\mathcal{S}} Y_j$ for (all) j in group $\mathcal{J}_{\boldsymbol{g}}$. Our next result shows that structural noncausality implies $G-$non-causality.

PROPOSITION 3.3. *Let A.1 - A.2 hold. Then structural noncausality* $(D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}, \forall \boldsymbol{g} \in \mathcal{G})$ *implies that D does not G−cause Y w.r.t. X, i.e.,* $D \perp Y \mid X$.

This result is intuitive: when structural effects are well defined and identifiable, $G-$non-causality follows from structural noncausality. Holland (1986, p. 958) gives a similar result for the random experiment case. Our Proposition 3.3 justifies using tests of $G-$noncausality to test structural noncausality: if we reject $G-$noncausality, we must reject structural noncausality.

As WL show, structural causality does not imply $G-$causality. Without further assumptions, the concepts are not equivalent. Nevertheless, if, as is commonly assumed in the literature (see WL for discussion), r obeys separability between observable causes D and unobservable causes U, or r obeys a specific form of monotonicity in U, then, with suitable regularity, structural causality does imply $G-$causality. We have

THEOREM 3.4. *Given A.1 - A.2, suppose* $\{Y_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ *is regular. Suppose further that for all* $\boldsymbol{g} \in \mathcal{G}$, *either* (i) *or* (ii) *holds:*

(i) *For all* $j \in \mathcal{J}_{\boldsymbol{g}}$, *for unknown measurable functions* r_1 *and* r_2,

$$Y_j = r_1\left(D_j, Z_j; b_j\right) + r_2\left(Z_j, U_j; b_j\right); \tag{3.3}$$

(ii) *For all* $j \in \mathcal{J}_{\boldsymbol{g}}$, *for* $\ell = 1, ..., k_y$, $r_\ell(d, z, u; b_j) = r_{0,\ell}(d, z, u_\ell; b_j)$ *for scalar* u_ℓ, *where* $r_{0,\ell}(d, z, \cdot; b_j)$ *is strictly monotone increasing for each admissible* (d, z), *and* $F_{j,\ell}(y_\ell \mid d, x) := P[Y_{j,\ell} \leq y_\ell \mid D_j = d, X_j = x]$ *is strictly monotone in* y_ℓ *for each admissible* (d, x).

Then structural causality (for some $\boldsymbol{g} \in \mathcal{G}$, $D_j \Rightarrow_{\mathcal{S}} Y_j$ *for (all) j in* $\mathcal{J}_{\boldsymbol{g}}$*) implies that D G−causes Y w.r.t. X, i.e.,* $D \not\perp Y \mid X$.

Kasy (2011) gives a discussion of structures satisfying (ii). Even without separability or monotonicity, there is an equivalence between $G-$causality and a stronger notion of structural causality that handles certain exceptional cases where the causal structure and the conditional distribution of (U_j, D_j) given X_j interact in just the right way to hide the structural causality. WL call this stronger notion *structural causality with positive probability* and provide discussions. We let $\mathbb{Y}_j := \text{supp}(Y_j)$ and $\mathbb{X}_j := \text{supp}(X_j)$.

DEFINITION 3.5 (Structural Causality with Positive Probability). Suppose A.1 holds, and let j be given. Suppose that for each $y \in \mathbb{Y}_j$, there exists a measurable function $f_{j,y} : \mathbb{X}_j \to [0, 1]$ such that

$$\int 1\left\{r\left(D_j, Z_j, u; b_j\right) < y\right\} \, dF_j\left(u \mid X_j\right) = f_{j,y}\left(X_j\right) \quad a.s., \tag{3.4}$$

where $F_j(u \mid X_j) := E[1\{U_j \leq u\} \mid X_j]$. Then D_j *does not structurally cause* Y_j *almost surely* $(a.s.)$ *w.r.t.* X_j $(D_j \not\Rightarrow_{\mathcal{S}(X_j)} Y_j)$. Otherwise, D_j *structurally causes* Y_j *with positive probability* $(w.p.p.)$ *w.r.t.* X_j $\left(D_j \Rightarrow_{\mathcal{S}(X_j)} Y_j\right)$.

By the definition of a group, the same structural causality $w.p.p.$ relations hold for all j in a given group. We thus write $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}(X_{\boldsymbol{g}})} Y_{\boldsymbol{g}}$ when $D_j \not\Rightarrow_{\mathcal{S}(X_j)} Y_j$ holds for all j in group $\mathcal{J}_{\boldsymbol{g}}$.

Below is an example in WL in which D_j structurally causes Y_j, while D_j does not structurally cause Y_j $a.s$. Consider the structural equation r

$$Y_j = r\left(D_j, U_j\right) = \frac{D_j}{\sqrt{D_j^2+1}} U_{j1} + \frac{1}{\sqrt{D_j^2+1}} U_{j2},$$

where $U_j \equiv (U_{j1}, U_{j2})$ and D_j are all $N(0,1)$ random variables and D_j, U_{1j}, and U_{2j} are mutually independent. For simplicity, there is no b_j, Z_j, or X_j. It is clear that D_j structurally causes Y_j here. Nevertheless, D_j does not structurally cause Y_j $a.s$. To see this, note that Y_j and D_j are independent. Thus, the LHS of eq.(3.4) becomes

$$\begin{aligned}
&\int 1\left\{r\left(D_j, u_1, u_2\right) < y\right\} dF_j\left(u_1, u_2\right) \\
&\quad = \int 1\left\{r\left(D_j, u_1, u_2\right) < y\right\} dF_j\left(u_1, u_2 | D_j\right) = \Pr\left[Y_j < y | D_j\right] = \Phi(y),
\end{aligned}$$

where $F_j(\cdot)$ and $F_j\left(\cdot | D_j\right)$ denote the CDF of U_j and the conditional CDF of U_j given D_j, and Φ is the standard normal CDF. Thus the RHS does not depend on D_j, i.e., D_j does not structurally cause Y_j $a.s$.

Theorem 3.5 below gives a result linking $G-$causality and structural causality $a.s$.

THEOREM 3.5. *Suppose A.1 - A.2 hold. Suppose also that* $\{Y_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}, \boldsymbol{g} \in \mathcal{G}\}$ *is regular. If* D_j *structurally causes* Y_j $w.p.p.$ *w.r.t.* X_j (*for some* $\boldsymbol{g} \in \mathcal{G}$, $D_j \Rightarrow_{\mathcal{S}(X_j)} Y_j$ *for all* j *in* $\mathcal{J}_{\boldsymbol{g}}$), *then* D $G-$*causes* Y *w.r.t.* X, *i.e.,* $D \not\perp Y \mid X$.

Together, Proposition 3.3 and Theorem 3.5 establish that cross-section $G-$causality and structural causality are essentially equivalent under the key conditional exogeneity assumption. We say "essentially" as Theorem 3.5 rules out the two ways that $G-$non-causality can mask structural causality. The first arises from subtle interactions between the causal structure and the conditional distribution of (U_j, D_j) given X_j. The second arises under heterogeneity, from delicate cancellations among the conditional distributions of Y_j given D_j and X_j across groups. These exceptional possibilities only trivially mitigate the power of tests for $G-$causality as tests for structural causality. Errors of interpretation and inference need not result, as long as these exceptions are recognized.

3.5. Identification and identification failure

To gain further insight into the relation between $G-$causality and effect identification, we now undertake a deeper analysis of $G-$causality in mean. For this,

let $q_{\boldsymbol{g}}(d,x) := f_{\boldsymbol{g}}(d,x)/\sum_{\boldsymbol{h}\in\mathcal{G}}\left(p_{\boldsymbol{h}} f_{\boldsymbol{h}}(d,x)\right)$. Under A.1, the law of iterated expectations gives

$$E\left(Y \mid D=d, X=x\right) = \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, q_{\boldsymbol{g}}(d,x) \int r(d,z,u;b_{\boldsymbol{g}})\, dF_{\boldsymbol{g}}(u \mid d,x).$$

For concreteness, consider identifying average marginal effects for a continuous treatment d. To allow both discrete and continuous U_g, let $\nu_{\boldsymbol{g}}(u \mid d,x)$ define a $\sigma-$finite measure dominating that defined by $F_{\boldsymbol{g}}(u \mid d,x)$ and let $f_{\boldsymbol{g}}(u \mid d,x)$ define the associated Radon-Nikodym density for $U_{\boldsymbol{g}}$ given $(D_{\boldsymbol{g}}, X_{\boldsymbol{g}})$. Assuming differentiability and mild regularity, derivations parallel to those of White and Chalak (2013) give

$$\begin{aligned}
&\frac{\partial}{\partial d} E(Y \mid D=d, X=x) \\
&\qquad = \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, q_{\boldsymbol{g}}(d,x) \int \frac{\partial}{\partial d} r(d,z,u;b_{\boldsymbol{g}}) f_{\boldsymbol{g}}(u \mid d,x)\, d\nu_{\boldsymbol{g}}(u \mid d,x) \\
&\qquad\quad + \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, q_{\boldsymbol{g}}(d,x) \int [r(d,z,u;b_{\boldsymbol{g}})\, \frac{\partial}{\partial d} \ln f_{\boldsymbol{g}}(u \mid d,x)] \\
&\qquad\quad \times f_{\boldsymbol{g}}(u \mid d,x)\, d\nu_{\boldsymbol{g}}(u \mid d,x) \\
&\qquad\quad + \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, \frac{\partial}{\partial d} q_{\boldsymbol{g}}(d,x) \int r(d,z,u;b_{\boldsymbol{g}})\, f_{\boldsymbol{g}}(u \mid d,,x) d\nu_{\boldsymbol{g}}(u \mid d,x).
\end{aligned}$$

After some manipulation, this regression derivative can be written as

$$\frac{\partial}{\partial d} E\left(Y \mid D=d, X=x\right) = \bar{\beta}(d,x) + \delta_1(d,x) + \delta_2(d,x). \tag{3.5}$$

The first of the three terms on the right is a weighted average marginal effect,

$$\begin{aligned}
\bar{\beta}(d,x) &:= \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, q_{\boldsymbol{g}}(d,x)\, \beta_{\boldsymbol{g}}(d,x), \text{ where } \beta_{\boldsymbol{g}}(d,x) \\
&:= E\left[(\partial/\partial d) r(D_{\boldsymbol{g}}, Z_{\boldsymbol{g}}, U_{\boldsymbol{g}}; b_{\boldsymbol{g}}) \mid D_{\boldsymbol{g}}=d, X_{\boldsymbol{g}}=x\right]
\end{aligned}$$

is the covariate-conditioned average marginal effect of D_j on Y_j at $(D_j=d, X_j=x)$ for $j \in \mathcal{J}_{\boldsymbol{g}}$, $\boldsymbol{g}\in\mathcal{G}$. Thus, when the other two terms vanish, the regression derivative identifies the structurally meaningful weighted effect $\bar{\beta}(d,x)$; and, when $\bar{\beta}(d,x)$ is nonzero for (d,x) in a set of positive probability, we have $G-$causality in mean of D for Y w.r.t. X. When the other two terms do not vanish, we generally have identification failure, and there is no necessary link between $G-$causality and structural causality.

The second term is a pure endogeneity bias,

$$\delta_1(d,x) := \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, q_{\boldsymbol{g}}(d,x)\, E\left(\varepsilon_{\boldsymbol{g}}\, S_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}=d, X_{\boldsymbol{g}}=x\right),$$

where $\varepsilon_{\boldsymbol{g}} := Y_{\boldsymbol{g}} - E\left(Y_{\boldsymbol{g}} \mid D_{\boldsymbol{g}} = d, X_{\boldsymbol{g}} = x\right)$ is the regression residual for the given group and $S_{\boldsymbol{g}} := (\partial/\partial d) \ln f_{\boldsymbol{g}}(U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}})$ is the *exogeneity score* of White and Chalak (2013). Observe that $E(\varepsilon_{\boldsymbol{g}} S_{\boldsymbol{g}} \mid D_{\boldsymbol{g}} = d, X_{\boldsymbol{g}} = x)$ represents a specific form of omitted variable bias, where the exogeneity score is the omitted variable. A sufficient (but not necessary) condition for this bias to vanish is A.2(ii), $D_{\boldsymbol{g}} \perp U_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$ for all $\boldsymbol{g} \in \mathcal{G}$.

The third term is a pure heterogeneity bias,

$$\delta_2(d,x) := \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}} \frac{\partial}{\partial d} q_{\boldsymbol{g}}(d,x)\, E\left(Y_{\boldsymbol{g}} \mid D_{\boldsymbol{g}} = d, X_{\boldsymbol{g}} = x\right),$$

due to heterogeneity of $f_{\boldsymbol{g}}(d \mid x)$ across members of $\mathcal{G}$.[8] A sufficient (but not necessary) condition for this to vanish is A.2(i), $D_{\boldsymbol{g}} \mid X_{\boldsymbol{g}} \sim D \mid X$ for all $\boldsymbol{g} \in \mathcal{G}$.

Thus, A.2 ensures that the regression derivative identifies the weighted effect

$$\bar{\beta}(d,x) = \bar{\beta}^*(d,x) := \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, q^*_{\boldsymbol{g}}(x)\, \beta^*_{\boldsymbol{g}}(d,x),$$

where $q^*_{\boldsymbol{g}}(x) := f_{\boldsymbol{g}}(x) / \sum_{\boldsymbol{h}} (p_{\boldsymbol{h}} f_{\boldsymbol{h}}(x))$ and $\beta^*_{\boldsymbol{g}}(d,x) := \int (\partial/\partial d) r(d,z,u;b_{\boldsymbol{g}})\, dF_{\boldsymbol{g}}(u \mid x)$. This identification provides the link between $G-$causality in mean and structural causality.

Although A.2 is not necessary for effect identification, cases where identification holds in the absence of A.2 are quite special, analogous to failures of regularity. Thus, for practical purposes, A.2 can be viewed as playing the key role in identifying effects of interest and thereby linking $G-$causality and structural causality.

Further, as Proposition 6.1 and Corollary 6.2 of WL show, in the absence of structural causality, $G-$causality is essentially equivalent to exogeneity failure. Here, this is reflected in the fact that when $\beta_{\boldsymbol{g}}(d,x)$ vanishes for all $\boldsymbol{g}$, A.1 and mild regularity conditions give

$$\frac{\partial}{\partial d} E\left(Y \mid D = d, X = x\right) = \delta_1(d,x) + \delta_2(d,x).$$

Thus, with structural noncausality *a.s.*, $G-$causality in mean implies the failure of A.2(i), A.2(ii), or both; conversely, failure of A.2 essentially ensures $G-$causality in mean.

3.6. Testing for $G-$causality and structural causality in cross-section data

As just seen, the hypothesis of $G-$non-causality, and thus of structural noncausality under the conditional exogeneity assumption, is a specific conditional independence. In the literature, there are many conditional independence tests that apply to IID data (e.g., see Delgado and González-Manteiga, 2001; Fernandes and

Flores, 2001; Su and White, 2007, 2008, 2014; Song, 2009; Linton and Gozalo, 2014; Huang, Sun, and White, 2016). These methods can be computationally challenging, as they are nonparametric. Angrist and Kuersteiner (2004, 2011) develop a semi-parametric test for conditional independence. Based on the fact that $Y \perp D \mid X$ implies $\psi_1(Y) \perp \psi_2(D) \mid X$ for any measurable vector functions (ψ_1, ψ_2), WL propose several convenient regression-based tests.

4. STRUCTURAL CAUSALITY AND $G-$CAUSALITY IN PANEL DATA

The literature contains considerable discussion about $G-$causality in the panel data setting. Nevertheless, the focus is mainly on linear or parametric models. For example, Chamberlain (1984) discusses $G-$causality "conditional on unobservables ['fixed effect']" in a linear model and a logit model. Holtz-Eakin, Newey, and Rosen (1988) consider a traditional linear vector autoregression (VAR) and use GMM to test $G-$causality. Nair-Reichert and Weinhold (2001) discuss $G-$causality in a dynamic mixed fixed- and random-coefficients linear model allowing heterogeneity across individuals. Dumitrescu and Hurlin (2012) test for $G-$causality in a linear dynamic panel model with fixed coefficients that vary across individuals. There is also a growing literature on nonlinear and nonseparable panel models (see, e.g., Hoderlein and White (2012) and the references therein). We are not aware that $G-$causality has been discussed in such models so far.

We emphasize that in panel data, $G-$causality is also entirely based on predictability and has no causal interpretation. In this section, we link $G-$causality and structural causality in general panel structures. Since the static panel case is similar to the cross-sectional case, we focus on a dynamic data generating process. To simplify notation, for given $\ell_y \in \mathbb{N}^+$, we let $\boldsymbol{Y}_{j,t-1}$ denote the finite history $\boldsymbol{Y}_{j,t-1} := (Y_{j,t-\ell_y}, ..., Y_{j,t-1})$.

Assumption B.1 (Panel Population DGP)**.** Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, let $N \in \mathbb{N}^+$, and let $\ell_y \in \mathbb{N}^+$. For the members of a population $j \in \{1, ..., N\}$, let the random vectors $Y_{j,t}$, $t = 1, 2, ...$, be structurally determined by a triangular system as

$$Y_{j,t} = r\left(\boldsymbol{Y}_{j,t-1}, D_{j,t}, Z_{j,t}, U_{j,t}; b_{j,t}\right), \tag{4.1}$$

where r is an unknown measurable $k_y \times 1$ function, $k_y \in \mathbb{N}^+$; $Y_{j,\tau}$ $(\tau = 1 - \ell_y, ..., 0)$, $D_{j,t}(k_d \times 1, k_d \in \mathbb{N}^+)$, $Z_{j,t}$ $(k_z \times 1, k_z \in \mathbb{N})$, and $U_{j,t}$ $(k_u \times 1, k_u \in \bar{\mathbb{N}})$ are vectors of nondegenerate random variables on $(\Omega, \mathcal{F}, P)$; and $b_{j,t}$ is a nonrandom real vector of dimension $k_b \in \bar{\mathbb{N}}$. Suppose also that $W_{j,t}$, $t = 1, 2, ...$, are random vectors on $(\Omega, \mathcal{F}, P)$ with dimension $k_w \in \mathbb{N}$. The triangular structure is such that for $t = 1, 2, ...$, neither $\{Y_{j,s}\}_{s=1-\ell_y}^{\infty}$ nor $\{D_{j,s}\}_{s=1}^{\infty}$ structurally determine $W_{j,t}$; $\{Y_{j,s}\}_{s=1-\ell_y}^{\infty}$ does not structurally determine $D_{j,t}$, $Z_{j,t}$, or $U_{j,t}$; and $\{D_{j,s}\}_{s=1}^{\infty}$ does not structurally determine $Z_{j,t}$ or $U_{j,t}$.

The interpretation of structural eq.(4.1) is the same as that for eq.(3.1), except that here we have multiple time periods $t = 1, 2, ...$ The elements of $\left(Y_{j,t}, D_{j,t}, Z_{j,t}, W_{j,t}, U_{j,t}\right)$ and $b_{j,t}$'s can contain both time-varying and time-invariant elements. For now, we do not distinguish these. The elements of $(D_{j,t}, Z_{j,t}, W_{j,t}, U_{j,t})$ can contain lags of underlying variables. By permitting only finite histories $\boldsymbol{Y}_{j,t-1}$, $D_{j,t}$, $Z_{j,t}$, we restrict attention to Markov-type data generating processes; this greatly simplifies the analysis and corresponds to the structures usually considered in practice.

As in the cross-section case, we can have both structural heterogeneity and distributional heterogeneity. We specify the relevant structural and distributional heterogeneities below.

We again impose random sampling from the population. For simplicity, we assume all time periods can be observed for every individual, so we observe a balanced panel. Similar to the cross-section case, we write

$$Y_t = r\left(\boldsymbol{Y}_{t-1}, D_t, Z_t, U_t; B_t\right),\ t = 1, ..., T,$$

by suppressing the cross-sectional subscript i. We assume that we observe data (Y_t, D_t, Z_t, W_t), $t = 1, ..., T$, and relevant lags prior to $t = 1$.

4.1. Linking structural causality and $G-$causality in dynamic panels

We first define structural causality for the dynamic panel structure. Let $\mathbb{D}_{j,t}$ be the support of $D_{j,t}$.

DEFINITION 4.1 (Structural Causality in Dynamic Panels). Let j and t be given. If the function $r(\boldsymbol{y}_{j,t-1}, \cdot, z_{j,t}, u_{j,t}; b_{j,t}) : \mathbb{D}_{j,t} \to \mathbb{R}^{k_y}$ is constant on $\mathbb{D}_{j,t}$ for all admissible $\boldsymbol{y}_{j,t-1}$, $z_{j,t}$ and $u_{j,t}$, then $D_{j,t}$ ***does not structurally cause*** $Y_{j,t}$, and we write $D_{j,t} \not\Rightarrow_{\mathcal{S}} Y_{j,t}$. Otherwise, $D_{j,t}$ ***structurally causes*** $Y_{j,t}$, and we write $D_{j,t} \Rightarrow_{\mathcal{S}} Y_{j,t}$.

We let $X_{j,t} := (Z'_{j,t}, W'_{j,t})'$ denote the covariates at time t. WL consider retrospective conditional exogeneity and retrospective $G-$causality (White and Kennedy, 2009). We also allow this. For this, we let $\boldsymbol{X}_{j,t}$ denote a covariate history that may contain $X_{j,t}$ and lags or leads of $X_{j,t}$, as in Wooldridge (2005, p. 41) or WL; i.e., for some $m \in \mathbb{N}^+$,

$$\boldsymbol{X}_{j,t} := \mathbb{S}_{\boldsymbol{X}}\left(X_{j,t-m}, ..., X_{j,t}, ...X_{j,t+m}\right), \quad t = 1, 2, ...,$$

where $\mathbb{S}_{\boldsymbol{X}}$ is a given selection matrix. We assume $\left\{X_{j,t}, ..., X_{j,t-m}\right\} \subseteq \boldsymbol{X}_{j,t}$ and allow $\boldsymbol{X}_{j,t}$ also to contain leads of $X_{j,t}$. When $\mathbf{X}_{j,t}$ contains leads, we have the retrospective case, as in WL. We denote randomly sampled covariates $\boldsymbol{X}_t$. For simplicity, we also assume that all needed elements of $\boldsymbol{X}_t$ are observable. Temporal precedence plays an important role in panel data. In particular, compared with the cross-sectional case, we now have a richer set of covariates (including both leads and lags) to choose from.

To provide a link between structural causality and $G-$causality for dynamic panels, we condition on not only covariates $\boldsymbol{X}_t$ but also lags of Y_t. For this, we denote $\tilde{\boldsymbol{Y}}_{j,t-1} := (Y_{j,t-m_y},...,Y_{j,t-1})$, where $m_y \in \mathbb{N}^+$ and $m_y \geq \ell_y$; i.e., $\{\boldsymbol{Y}_{j,t-1}\} \subseteq \{\tilde{\boldsymbol{Y}}_{j,t-1}\}$. Similarly, we let $\tilde{\boldsymbol{Y}}_{t-1}$ be the randomly sampled $\tilde{\boldsymbol{Y}}_{j,t-1}$.

Similar to the cross-section case, for each t, we define population group $\mathcal{J}_{\boldsymbol{g},t}$, $\boldsymbol{g} \in \{1,...,\Gamma_t\}$, as a collection of population units having identical $b_{j,t}$ and identical distribution of $(U_{j,t}, D_{j,t}) \mid \boldsymbol{X}_{j,t}, \tilde{\boldsymbol{Y}}_{j,t-1}$. We also let $\mathcal{G}_t := \{1,...,\Gamma_t\}$. For given t, we write $D_{\boldsymbol{g},t} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g},t}$ to denote $D_{j,t} \not\Rightarrow_{\mathcal{S}} Y_{j,t}$ for all j in group $\mathcal{J}_{\boldsymbol{g},t}$, $\boldsymbol{g} \in \mathcal{G}_t$.

Next, we impose an exogeneity assumption analogous to A.2.

Assumption B.2 (Heterogeneous Dynamic Panel Exogeneity)**.** Let t be given. For all $\boldsymbol{g} \in \mathcal{G}_t$, (i) $D_{\boldsymbol{g},t} \mid \boldsymbol{X}_{\boldsymbol{g},t}, \tilde{\boldsymbol{Y}}_{\boldsymbol{g},t-1} \sim D_t \mid \boldsymbol{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$; and (ii) $D_{\boldsymbol{g},t} \perp U_{\boldsymbol{g},t} \mid \boldsymbol{X}_{\boldsymbol{g},t}, \tilde{\boldsymbol{Y}}_{\boldsymbol{g},t-1}$.

Here, as before for all $\boldsymbol{g} \in \mathcal{G}_t$, the conditional distributions of $D_{\boldsymbol{g},t}$ given $(\boldsymbol{X}_{\boldsymbol{g},t}, \tilde{\boldsymbol{Y}}_{\boldsymbol{g},t-1})$ are assumed to be identical and we write the common distribution as $D_t \mid \boldsymbol{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$.

The analog of Theorem 3.1, relating population and sample conditional exogeneity, is

THEOREM 4.1. *Suppose B.1–B.2*(i) *hold, and let* t *be given. If B.2*(ii) *holds, then* $D_t \perp U_t \mid \boldsymbol{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$.

Proposition 4.2 provides a link between structural noncausality and $G-$noncausality.

PROPOSITION 4.2. *Suppose that B.1 and B.2 hold. Let* t *be given. Then structural noncausality* $(D_{\boldsymbol{g},t} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g},t},\ \forall \boldsymbol{g} \in \mathcal{G}_t)$ *implies that* D_t *does not* $G-$*cause* Y_t *w.r.t.* $\boldsymbol{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$, *i.e.,* $D_t \perp Y_t \mid \boldsymbol{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$.

We can also define structural causality with positive probability for dynamic structures and show that structural causality and $G-$causality are "essentially" equivalent under the key conditional exogeneity assumption. For brevity, we omit the details.

So far, we have not carefully distinguished between time-varying and time-invariant elements of unobservable $b_{j,t}$ and $U_{j,t}$. Below, we briefly discuss the case where we allow time-invariant components of $U_{j,t}$ and the randomly sampled $b_{j,t}$ to be arbitrarily dependent or correlated with the cause of interest $D_{j,t}$. For this, we denote $U_{j,t} \equiv (\bar{U}_{j,t}, \bar{U}_{j,0})$, where $\bar{U}_{j,t}$ is time varying and $\bar{U}_{j,0}$ time-invariant. Similarly, we denote $b_{j,t} \equiv (\bar{b}_{j,t}, \bar{b}_{j,0})$. The randomly sampled $\bar{U}_{j,t}$, $\bar{U}_{j,0}$, $\bar{b}_{j,t}$, $\bar{b}_{j,0}$ are denoted as $\bar{U}_t$, $\bar{U}_0$, $\bar{B}_t$, $\bar{B}_0$, respectively. In principle, we want to remove the time-invariant $\bar{U}_{j,0}$ and $\bar{b}_{j,0}$, using, e.g., the first difference. For this, it is convenient to impose a separable assumption on the structural equation.[9] We also impose a conditional exogeneity assumption at the sample level.

Assumption B.3. Suppose that B.1 holds with

$$Y_{j,t} = r_1\left(D_{j,t}, \bar{U}_{j,0}; \bar{b}_{j,0}\right) + r_2\left(\mathbf{Y}_{j,t-1}, D_{j,t}, Z_{j,t}, \bar{U}_{j,t}; \bar{b}_{j,t}\right),\ j = 1, ..., N,\ t = 1, ...,$$

where r_1 and r_2 are two unknown $k_y \times 1$ measurable functions.

Assumption B.4. Suppose that for given t, $D_t \perp \left(\bar{U}_t, \bar{U}_{t-1}, \bar{B}_t, \bar{B}_{t-1}\right) \mid \mathbf{X}_t, \tilde{\mathbf{Y}}_{t-1}$.

B.4 is a sample-level assumption, which can be supported by a corresponding assumption at the heterogeneous population level. Note that in general, B.4 is a strong assumption, as we condition on $\tilde{\mathbf{Y}}_{t-1}$, which is a function of both D_{t-1} and $\bar{U}_{t-1}$. Nevertheless, this assumption is plausible under the null of structural noncausality, as $\tilde{\mathbf{Y}}_{j,t-1}$ is a constant function of $D_{j,t-1}$ under the null. One simple example is that $b_{j,t}$'s are constants over j and t and $\{D_{j,t}\}_{t=1}^{\infty} \perp \{\{Y_{j,t}\}_{t=1-\ell_y}^{0}, \{U_{j,t}\}_{t=1}^{\infty}, \{X_{j,t}\}_{t=1}^{\infty}\}$. Under the null of structural noncausality for all t, $\tilde{\mathbf{Y}}_{j,t-1}$ is a function of $\{\{Y_{j,s}\}_{s=1-\ell_y}^{0}, \{U_{j,s}\}_{s=1}^{t}, \{X_{j,s}\}_{s=1}^{t}\}$. Then it is easy to show that in this case, B.4 is satisfied. Certainly, it will be interesting to relax B.4, but we leave this for future research.

PROPOSITION 4.3. *Suppose that B.1–B.4 hold. Suppose that* $D_{\boldsymbol{g},t} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g},t}$ *and* $D_{\boldsymbol{g},t-1} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g},t-1}\ \forall \boldsymbol{g} \in \mathcal{G}_t \cup \mathcal{G}_{t-1}$. *Then,* $D_t \perp Y_t \mid \mathbf{X}_t, \tilde{\mathbf{Y}}_{t-1}$.

Proposition 4.3 suggests that when there are time-invariant components in the unobservables, we can test for structural noncausality by testing for $G-$noncausality $D_t \perp Y_t \mid \mathbf{X}_t, \tilde{\mathbf{Y}}_{t-1}$ under the conditional exogeneity assumption B.4.

4.2. Testing for $G-$causality and structural causality in panel data

As shown above, we can perform a $G-$non causality test to test for structural noncausality. Testing $G-$non-causality in panel data is also simply a conditional independence test. Here we focus on testing $D_t \perp Y_t \mid \mathbf{X}_t, \tilde{\mathbf{Y}}_{t-1}$.

First, assume that the joint distributions of (D_t, Y_t, X_t) are identical over t. In this case, we can pool the data and implement a conditional independence test as discussed in Section 3.6.

Second, suppose that the joint distributions of (D_t, Y_t, X_t) are different over time t. In this case, for each time period t, using the cross-section data, we can implement a conditional independence test for $D_t \perp Y_t \mid (\mathbf{X}_t, \tilde{\mathbf{Y}}_{t-1})$ and obtain a test statistic, say $\mathcal{W}_t$. Thus we have T test statistics, $\{\mathcal{W}_1, \mathcal{W}_2, ..., \mathcal{W}_T\}$ and each one can be used to test for structural noncausality for each t. We may also want to test the hypothesis that for all t, $D_{j,t}$ does not structurally cause $Y_{j,t}$. For this, we can construct our test statistic by taking the average of the T test statistics, i.e., $\mathcal{W} \equiv \frac{1}{T}\sum_{t=1}^{T} \mathcal{W}_t$, as in Im, Pesaran, and Shin (2003) and Dumitrescu and Hurlin (2012). Other forms of "averages" are possible and special care is needed to take into account the dependence structure over time.

5. CONCLUSION

This paper provides direct links between Granger causality and structural causality in cross-section and panel data. We extend Granger causality to cross-section and panel data and give a natural definition of structural causality in heterogeneous populations. We show that under the key conditional exogeneity assumption, Granger causality is essentially equivalent to structural causality in cross-section and panel data. Similar to the results in White and Lu (2010) for time-series data, our results here should enable researchers to avoid the misuse of Granger causality and to establish the desired structural causal relation.

NOTES

1. Strictly speaking, if the conditional distributions are parameterized or treated as infinite-dimensional parameters, then distributional heterogeneity may also be thought of as a special case of structural heterogeneity.

2. The classification is certainly overly simplistic. We ignore the early literature on the philosophy of causality (for a review, see, e.g., Holland, 1986 and Hoover, 2008). We also ignore the literature on causality discussed in other disciplines, notably in machine learning (see, e.g., Pearl, 2009, 2015, White and Chalak, 2009, and White, Chalak, and Lu, 2011). In particular, Heckman and Pinto (2015) compare Haavelmo's structural framework of causality with the framework of Directed Acyclic Graphs (DAG) studied in the Bayesian network.

3. Although U_j is typically called "unobservable," it is better to view its elements as variables that will be omitted from empirical analysis. This may be because they are unobservable; it may also be because the researcher has purposefully or inadvertantly neglected them.

4. For notational simplicity, we suppress the observation subscript i.

5. Using our notation, Holland considers the case where $D \perp (X, U)$, which implies our conditional exogeneity assumption $D \perp U \mid X$ below.

6. See, e.g., Altonji and Matzkin (2005), Hoderlein and Mammen (2007, 2009), Imbens and Newey (2009), White and Lu (2011), and White and Chalak (2013).

7. $ATT = E(Y(1) - Y(0) \mid D = 1) = E(U - 0 \mid D = 1) = E(U) = 0$.

8. The conditional density $f_{\boldsymbol{g}}(d \mid x)$ can also be interpreted as Imbens's (2000) *generalized propensity score*.

9. There is a recent literature which allows time-invariant unobservables to be arbitrarily correlated with causes of interest in the general nonseparable models (see, e.g., Evdokimov, 2010). We leave the problem of linking G-causality and structural causality in this general setting for future research.

REFERENCES

Altonji, J. & R. Matzkin (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73, 1053–1102.

Angrist, J., K. Graddy, & G. Imbens (2000) The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies* 67, 499–527.

Angrist, J., Ò. Jordà, & G. Kuersteiner (2013) Semiparametric estimates of monetary policy effects: string theory revisited. NBER *Working Paper* 19355.

Angrist, J. & G. Kuersteiner (2004) Semiparametric causality tests using the policy propensity score. NBER *Working Paper* No. 10975.

Angrist, J. & G. Kuersteiner (2011) Causal effects of monetary shocks: semiparametric conditional independence tests with a multinomial propensity score. *Review of Economics and Statistics* 93, 725–747.

Angrist, J. & J. Pischke (2009) *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bester, C.A. & C.B. Hansen (2016) Grouped effects estimators in fixed effects models. *Journal of Econometrics* 190, 197–208.

Bonhomme, S. & E. Manresa (2015) Grouped patterns of heterogeneity in panel data. *Econometrica* 83, 1147–1184.

Browning, M. & J.M. Carro (2007) Heterogeneity and microeconometrics modelling. In R. Blundell, W. K. Newey and T. Persson (Eds.), *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, Volume 3, pp. 45–74. Cambridge University Press.

Browning, M. & J.M. Carro (2010) Heterogeneity in dynamic discrete choice models. *Econometrics Journal* 13, 1–39.

Burda, M., M. Harding, & J. Hausman (2015) A Bayesian semi-parametric competing risk model with unobserved heterogeneity. *Journal of Applied Econometrics* 30, 353–376.

Chalak, K. & H. White (2011) An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics* 44, 1–51.

Chalak, K. & H. White (2012) Causality, conditional independence, and graphical separation in settable systems. *Neural Computation* 24, 1611–1668.

Chamberlain, G. (1982) The general equivalence of Granger and Sims causality. *Econometrica* 50, 569–581.

Chamberlain, G. (1984) Panel data. In Z. Griliches & M. Intriligator (eds.), *Handbook of Econometrics*, vol.2, ch. 22, pp.1247–1318. Elsevier.

Chesher, A. (2003) Identification in nonseparable models. *Econometrica* 71, 1405–1441.

Chesher, A. (2005) Nonparametric identification under discrete variation. *Econometrica* 73, 1525–1550.

Dawid, A.P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, *Series B* 41, 1–31.

Deb, P. & P.K. Trivedi (2013) Finite mixture for panels with fixed effects. *Journal of Econometric Methods* 2, 35–51.

Delgado, M. & W. González-Manteiga (2001) Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics* 29, 1469–1507.

Dumitrescu, E-I. & C. Hurlin (2012) Testing for Granger non-causality in heterogeneous panels. *Economic Modelling* 29, 1450–1460.

Evdokimov, K. (2010) Identification and estimation of a nonparametric panel data model with unobserved heterogeneity. *Working paper*, Department of Economics, Princeton University.

Florens, J.P. (2003) Some technical issues in defining causality. *Journal of Econometrics* 112, 127–128.

Florens, J.P. & D. Fougère (1996) Non-causality in continuous time. *Econometrica* 64, 1195–1212.

Florens, J.P. & M. Mouchart (1982) A note on non-causality. *Econometrica* 50, 583–591.

Granger, C.W.J. (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438.

Granger, C.W.J. (1980) Testing for causality, a personal viewpoint. *Journal of Economic Dynamics and Control* 2, 329–352.

Granger, C.W.J. (1988) Some recent developments in a concept of causality. *Journal of Econometrics* 39, 199–211.

Granger, C.W.J. & P. Newbold (1986) *Forecasting Economic Time Series*. Academic Press, 2nd edition.

Hausman, J. & T. Woutersen (2014) Estimating a semi-parametric duration model without specifying heterogeneity. *Journal of Econometrics* 178, 114–131.

Haavelmo, T. (1943) The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1–12.

Haavelmo, T. (1944) The probability approach in econometrics. *Econometrica* 12 (supp.), iii-vi, 1–115.

Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effect. *Econometrica* 66, 315–331.
Hahn, J. & R. Moon (2010) Panel data models with finite number of multiple equilibria. *Econometric Theory* 26, 863–881.
Heckman, J.J. (2000) Causal parameters and policy analysis in economics: a twentieth century retrospective. *Quarterly Journal of Economics* 115, 45–97.
Heckman, J.J. (2008) Econometric causality. *International Statistical Review* 76, 1–27.
Heckman, J.J. & R. Pinto (2015) Causal analysis after Haavelmo. *Econometric Theory* 31, 115–151.
Hirano, K., G. Imbens, & G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71, 1161–1189.
Hoderlein, S. & E. Mammen (2007) Identification of marginal effects in nonseparable models without monotonicity. *Econometrica* 75, 1513–1519.
Hoderlein, S. & E. Mammen (2009) Identification and estimation of local average derivatives in nonseparable models without monotonicity. *Econometrics Journal* 12, 1–25.
Hoderlein, S. & H. White (2012) Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics* 168, 300–314.
Holland, P. (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81, 945–970.
Holtz-Eakin, D., W. Newey, & H.S. Rosen (1988) Estimating vector autoregressions with panel data. *Econometrica* 56, 1371–1396.
Hoover, K. (2008) Causality in economics and econometrics. In S. Durlauf and L. Blume (eds.), *The New Palgrave Dictionary of Economics*, 2nd edition, Palgrave Macmillan.
Hsiao, C. (2014) *Analysis of Panel Data*. Cambridge University Press.
Huang, M., Y. Sun & H. White (2016) A flexible nonparametric test for conditional independence. *Econometric Theory*, forthcoming.
Hurwicz, L. (1950) Generalization of the concept of identification. In T.C. Koopmans (eds.), *Statistical Inference in Dynamic Economic Models*, pp. 238–257. John Wiley.
Im, K., H. Pesaran, & Y. Shin (2003) Testing for unit roots in heterogeneous panels. *Journal of Econometrics* 115, 53–74.
Imbens, G. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 706–710.
Imbens, G. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 86, 4–29.
Imbens, G. & W. Newey (2009) Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77, 1481–1512.
Imbens, G. & J. Wooldridge (2009) Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47, 5–86.
Kasy, M. (2011) Identification in triangular systems using control functions. *Econometric Theory* 27, 663–671.
Koopmans, T. (1950) *Statistical Inference in Dynamic Economic Models*. Cowles Commission Monograph No. 10, Wiley.
Kuersteiner, G. (2008) Granger–Sims causality. In S. Durlauf and L. Blume (Eds.), *The New Palgrave Dictionary of Economics*, 2nd edition, Palgrave Macmillan.
Lechner, M. (2001) Identification and estimation of causal effect of multiple treatments under conditional independence assumption. In M. Lechner, F. Pfeiller (eds), *Econometric Evaluations of Labour Market Policies*, Hildelberg, Physical, 43–58.
Lechner, M. (2011) The relation of different concepts of causality used in time series and microeconometrics. *Econometric Reviews* 30, 109–127.
Lin, C-C. & S. Ng (2012) Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods* 1, 42–55.
Lindsay, B. (1995) *Mixture Models: Theory, Geometry and Applications*. Hayward: Institute for Mathematical Statistics.

Linton, O. & P. Gozalo (2014) Testing conditional independence restrictions. *Econometric Reviews* 33, 523–552.

Lu, X. & L. Su (2014) Determining the number of groups in latent panel structures. *Working paper*, School of Economics, Singapore Management University.

Matzkin, R. (2003). Nonparametric estimation of nonadditive random functions. *Econometrica* 71, 1339–1375.

Matzkin, R. (2007). Nonparametric identification. In J.J. Heckman and E.E. Leamer (eds.), *The Handbook of Econometrics*, 6B. North-Holland.

Mclachlan, G. & K. Basford (1988) *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker.

Nair-Reichert, U. & D. Weinhold (2001) Causality tests for cross-country panels: a new look at FDI and economic growth in developing countries. *Oxford Bulletin of Economics and Statistics* 63, 153-171.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference* (2nd edition). Cambridge University Press.

Pearl, J. (2015) Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory* 31, 152–179.

Rosenbaum, P. & 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 non-randomized studies. *Journal of Educational Psychology* 66, 688–701.

Rubin, D. (2004) Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 31, 161–170.

Sarafidis, V. & N. Weber (2015) A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics* 77, 274–296.

Sims, C. (1972) Money, income and causality. *American Economic Review* 62, 540–552.

Song, K. (2009) Testing conditional independence via Rosenblatt transforms. *Annals of Statistics* 37, 4011–4015.

Strotz, R. & H. Wold (1960) Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica* 28, 417–427.

Su, L., Z. Shi, & P.C.B. Phillips (2014) Identifying latent structures in panel data. *Working paper*, Department of Economics, Yale University.

Su, L. & H. White (2007) A Consistent characteristic function-based test for conditional independence. *Journal of Econometrics* 141, 807–834.

Su, L. & H. White (2008) A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* 24, 829–864.

Su, L. & H. White (2014) Testing conditional independence via empirical likelihood. *Journal of Econometrics* 182, 27–44.

Sun, Y. (2005) Estimation and inference in panel structure models. *Working paper*, Department of Economics, UCSD.

White, H. & K. Chalak (2009) Settable systems: An extension of Pearl's causal model with optimization, equilibrium, and learning. *Journal of Machine Learning Research* 10, 1759–1799.

White, H. & K. Chalak (2010) Testing a conditional form of exogeneity. *Economics Letters* 109, 88–90.

White, H. & K. Chalak (2013) Identification and identification failure for treatment effects using structural systems. *Econometric Reviews* 32, 273–317.

White, H., K. Chalak, & X. Lu (2011) Linking Granger causality and the Pearl causal model with settable systems. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 12, 1–29.

White, H. & P. Kennedy (2009) Retrospective estimation of causal effects through time. In J. Castle & N. Shephard (eds.), *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry,* pp. 59–87. Oxford University Press.

White, H. & X. Lu (2010) Granger causality and dynamic structural systems. *Journal of Financial Econometrics* 8, 193–243.

White, H. & X. Lu (2011) Causal diagrams for treatment effect estimation with application to selection of efficient covariates. *Review of Economics and Statistics* 93, 1453–1459.

Wooldridge, J. (2005) Simple solutions to the initial conditions problem for dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* 20, 39–54.

Zellner, A. (1979) Causality and econometrics. In K. Brunner and A.H. Meltzer (eds), *Three Aspects of Policy and Policymaking*, *Carnegie-Rochester Conference Series*, Vol. 10, North-Holland, 9–54.

APPENDIX

Proof of Theorem 3.1. To show $D \perp U \mid X$, we establish that $f(u \mid d,x) = f(u \mid x)$ for all (u,d,x). We have

$$\begin{aligned} f(u \mid d,x) &= \frac{f(u,d,x)}{f(d,x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u,d,x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(d,x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d,x)\, f_{\boldsymbol{g}}(d,x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(d,x)} \\ &= \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d,x)\, f(d \mid x)\, f_{\boldsymbol{g}}(x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f(d \mid x)\, f_{\boldsymbol{g}}(x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid x)\, f_{\boldsymbol{g}}(x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(x)} \\ &= \frac{f(u,x)}{f(x)} = f(u \mid x), \end{aligned}$$

where the first equality in the second line holds since A.2(i) (i.e., $D_{\boldsymbol{g}} \mid X_{\boldsymbol{g}} \sim D \mid X$) implies $f_{\boldsymbol{g}}(d,x) = f(d \mid x)\, f_{\boldsymbol{g}}(x)$, say, and the second holds by A.2(ii), $f_{\boldsymbol{g}}(u \mid d,x) = f_{\boldsymbol{g}}(u \mid x)$. ■

Proof of Theorem 3.2. Note that

$$f(u \mid x) = \frac{f(u,x)}{f(x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid x)\, f_{\boldsymbol{g}}(x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(x)}.$$

We also have

$$\begin{aligned} f(u \mid d,x) &= \frac{f(u,d,x)}{f(d,x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u,d,x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(d,x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d,x)\, f_{\boldsymbol{g}}(d,x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(d,x)} \\ &= \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d,x)\, f(d \mid x) f_{\boldsymbol{g}}(x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f(d \mid x) f_{\boldsymbol{g}}(x)} = \frac{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d,x)\, f_{\boldsymbol{g}}(x)}{\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(x)}. \end{aligned}$$

The first equality in the second line follows from A.2(i). Then $f(u \mid d,x) = f(u \mid x)$ for all (u,d,x) if and only if

$$\sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid d,x)\, f_{\boldsymbol{g}}(x) = \sum_{\boldsymbol{g}\in\mathcal{G}} p_{\boldsymbol{g}}\, f_{\boldsymbol{g}}(u \mid x) f_{\boldsymbol{g}}(x) \quad \text{for all } (u,d,x).$$

By assumption, $\{(U_{\boldsymbol{g}} \mid D_{\boldsymbol{g}}, X_{\boldsymbol{g}}), \boldsymbol{g} \in \mathcal{G}\}$ is regular. It follows immediately that for all $\boldsymbol{g} \in \mathcal{G}$, $D_{\boldsymbol{g}} \perp U_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$. ■

Proof of Proposition 3.3. Take any $\boldsymbol{g} \in \mathcal{G}$ and let j belong to group $\mathcal{J}_{\boldsymbol{g}}$, $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}$ means that $Y_j = \tilde{r}(Z_j, U_j; b_j)$. Thus $U_j \perp D_j \mid X_j$ implies that $Y_j \perp D_j \mid X_j$ by Dawid (1979, Lemmas 4.1 and 4.2). This holds for all $\boldsymbol{g} \in \mathcal{G}$. Applying Theorem 3.1 gives $D \perp Y \mid X$. ■

Proof of Theorem 3.4. We prove this by showing that $Y \perp D \mid X$ implies that $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}$ for all $\boldsymbol{g} \in \mathcal{G}$. Given the assumed regularity, Theorem 3.2 ensures that $Y \perp D \mid X$ implies that $Y_{\boldsymbol{g}} \perp D_{\boldsymbol{g}} \mid X_{\boldsymbol{g}}$ for all $\boldsymbol{g} \in \mathcal{G}$.

Given (i), eq.(3.3) and A.2(ii) imply that for all j in group $\mathcal{J}_{\boldsymbol{g}}$

$$\begin{aligned} E\left(Y_j | X_j = x, D_j = d\right) &= r_1\left(d, z; b_j\right) + E\left[r_2\left(z, U_j; b_j\right) \mid D_j = d, X_j = x\right] \\ &= r_1\left(d, z; b_j\right) + E\left[r_2\left(z, U_j; b_j\right) \mid X_j = x\right]. \end{aligned}$$

Clearly, $r_1\left(d, z; b_j\right)$ is constant in d if and only if $E\left(Y_j | D_j = d, X_j = x\right)$ is constant in d. $Y_j \perp D_j \mid X_j$ implies that $E\left(Y_j | D_j = d, X_j = x\right)$ is constant in d, which then implies that $r_1\left(d, z; b_j\right)$ is constant in d. Thus $D_j \not\Rightarrow_{\mathcal{S}} Y_j$, so $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}$.

Given (ii), let $k_y = 1$ without loss of generality. Then by A.2(ii) for all j in group $\mathcal{J}_{\boldsymbol{g}}$

$$\begin{aligned} F_{U_j}(u \mid x) \; &:= E\left[1\left\{U_j \le u\right\} \middle| X_j = x\right] \\ &= E\left[1\left\{r\left(d, z, U_j; b_j\right) \le r\left(d, z, u; b_j\right)\right\} \middle| X_j = x\right] \\ &= E\left[1\left\{r\left(D_j, Z_j, U_j; b_j\right) \le r\left(d, z, u; b_j\right)\right\} \middle| D_j = d, X_j = x\right] \\ &= E\left[1\left\{Y_j \le r\left(d, z, u; b_j\right)\right\} \middle| D_j = d, X_j = x\right] \\ &=: F_j\left(r\left(d, z, u; b_j\right) \mid d, x\right). \end{aligned}$$

By strict monotonicity of $F_j(y \mid d, x)$ in y,

$$r\left(d, z, u; b_j\right) = F_j^{-1}\left(F_{U_j}(u \mid x) \mid d, x\right).$$

Now $Y_j \perp D_j \mid X_j$ implies that $F_j^{-1}\left(F_{U_j}(u \mid x) \mid d, x\right)$ is constant in d, so $r\left(d, z, u; b_j\right)$ is also a constant function in d, i.e., $D_j \not\Rightarrow_{\mathcal{S}} Y_j$, so $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}$.

As either (i) or (ii) holds for each $\boldsymbol{g} \in \mathcal{G}$, we have $D_{\boldsymbol{g}} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g}}$ for all $\boldsymbol{g} \in \mathcal{G}$. ■

Proof of Theorem 3.5. To show that A.1 - A.2 and the assumed regularity for $Y_{\boldsymbol{g}}$, together with $D_{\boldsymbol{g}} \Rightarrow_{\mathcal{S}(X_{\boldsymbol{g}})} Y_{\boldsymbol{g}}$ for some $\boldsymbol{g} \in \mathcal{G}$ imply $D \not\perp Y \mid X$, we let j in group $\mathcal{J}_{\boldsymbol{g}}$, $\boldsymbol{g} \in \mathcal{G}$, be such that $D_j \Rightarrow_{\mathcal{S}(X_j)} Y_j$. A.2$(ii)$ ensures that $D_j \perp U_j \mid X_j$, so

$$\begin{aligned} P[Y_j \le y \mid D_j = d, X_j = x] &= \int 1\left\{r(d, z, u; b_j) \le y\right\} dF_j(u \mid d, x) \\ &= \int 1\left\{r(d, z, u; b_j) \le y\right\} dF_j(u \mid x). \end{aligned}$$

By assumption, $D_j \Rightarrow_{\mathcal{S}(X_j)} Y_j$, so there exists $y^* \in \text{supp}(Y_j)$ such that there is no measurable mapping f_{j,y^*} for which $\int 1\{r(d, z, u; b_j) \le y^*\} dF_j(u \mid d, x) = f_{j,y^*}(x)$ $a.e.\text{-}x$. Specifically, this rules out the possibility that

$$\int 1\left\{r(d, z, u; b_j) \le y^*\right\} dF_j(u \mid X_j) = f_{j,y^*}(X_j) \equiv P[Y_j \le y^* \mid X_j] \quad a.s.$$

Thus, there exists $y^* \in \text{supp}(Y_j)$ such that

$$P\left[Y_j \le y^* \mid D_j = d, X_j = x\right] \neq P\left[Y_j \le y^* \mid X_j = x\right]$$

for x in a set of positive probability, so $Y_j \not\perp D_j \mid X_j$. Given the regularity assumed for $Y_{\boldsymbol{g}}$, Theorem 3.2 (for $Y_{\boldsymbol{g}}$) gives $D \not\perp Y \mid X$. ■

Proof of Theorem 4.1. The proof is analogous to that of Theorem 3.1. ■

Proof of Proposition 4.2. The proof is analogous to that of Proposition 3.3. ■

Proof of Proposition 4.3. Under B.3, $D_{\boldsymbol{g},t} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g},t}$ and $D_{\boldsymbol{g},t-1} \not\Rightarrow_{\mathcal{S}} Y_{\boldsymbol{g},t-1}$ $\forall \boldsymbol{g} \in \mathcal{G}_t \cup \mathcal{G}_{t-1}$ imply that for $j \in \mathcal{J}_{\boldsymbol{g}}$, there exist two measurable functions $\tilde{r}_1$ and $\tilde{r}_2$ such that

$$Y_{j,t} = \tilde{r}_1(\bar{U}_{j,0}, \bar{b}_{j,0}) + \tilde{r}_2\left(\boldsymbol{Y}_{j,t-1}, Z_{j,t}, \bar{U}_{j,t}, \bar{b}_{j,t}\right),\ j = 1, ..., N,\ t = 1, ...$$

Hence,

$$Y_{j,t} - Y_{j,t-1} = \tilde{r}_2\left(\boldsymbol{Y}_{j,t-1}, Z_{j,t}, \bar{U}_{j,t}; \bar{b}_{j,t}\right) - \tilde{r}_2\left(\boldsymbol{Y}_{j,t-2}, Z_{j,t-1}, \bar{U}_{j,t-1}; \bar{b}_{j,t-1}\right).$$

At the sample level, this means that

$$Y_t - Y_{t-1} = \tilde{r}_2\left(\boldsymbol{Y}_{t-1}, Z_t, \bar{U}_t; \bar{B}_t\right) - \tilde{r}_2\left(\boldsymbol{Y}_{t-2}, Z_{t-1}, \bar{U}_{t-1}; \bar{B}_{t-1}\right).$$

Then B.4 implies that

$$D_t \perp (\mathbf{X}_t, \tilde{\boldsymbol{Y}}_{t-1}, \bar{U}_t, \bar{U}_{t-1}, \bar{B}_t, \bar{B}_{t-1}) \mid \mathbf{X}_t, \tilde{\boldsymbol{Y}}_{t-1},$$

which further implies that $D_t \perp (Y_t - Y_{t-1}) \mid \mathbf{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$ by Dawid (1979, Lemmas 4.1 and 4.2), as $(Y_t - Y_{t-1})$ is a function of $(\mathbf{X}_t, \tilde{\boldsymbol{Y}}_{t-1}, \bar{U}_t, \bar{U}_{t-1}, \bar{B}_t, \bar{B}_{t-1})$. Then we have $D_t \perp Y_t \mid \mathbf{X}_t, \tilde{\boldsymbol{Y}}_{t-1}$ as Y_{t-1} is a component of $\tilde{\boldsymbol{Y}}_{t-1}$. ■