Identification and Estimation of Nonparametric Quantile Regressions with Endogeneity

Alexander Torgovitsky*

November 12, 2010

Job Market Paper

Abstract

I consider identification of nonparametric quantile regressions with endogenous regressors and an excluded instrumental variable. This model has an outcome equation that is both nonlinear and nonseparable in a latent variable which may be arbitrarily dependent with the regressors. This allows for general unobserved heterogeneity and selection on unobservables. I introduce a new assumption called conditional copula invariance (CCI) that restricts how the instrument affects the copula between the regressors and the latent variable. I show that CCI is sufficient for nonparametric identification given an extremely weak local rank condition. In particular, a binary instrument can satisfy the rank condition. This contrasts sharply with related work on nonparametric identification in nonseparable models, which has typically emphasized strong rank conditions such as large support or completeness. Furthermore, I show that CCI has a clear economic interpretation that can be justified by either a formal model or intuitive arguments. I use the identification result to develop a consistent nonparametric sieve extremum estimator. A Monte Carlo experiment shows that binary instruments perform comparably to continuous instruments in finite samples.

*Ph.D. candidate, Department of Economics, Yale University, alexander.torgovitsky@yale.edu. I thank my advisors, Donald Andrews, Xiaohong Chen and Edward Vytlacil, who have been gracious with their advice, support and feedback. This paper has benefited from discussions with Joseph Altonji, Lanier Benkard, Steven Berry, Kirill Evdokimov, Philip Haile, James Heckman, Kaisuke Hirano, Yuichi Kitamura, Sokbae (Simon) Lee and Oliver Linton. I thank participants of the 2010 Cowles Foundation Summer Conference, the 2010 World Congress of the Econometric Society and of econometrics workshops and seminars at Yale. While writing this paper I was generously supported by a Carl Arvid Anderson fellowship from the Cowles Foundation.

JEL classification: C14; C20; C51

Keywords: Nonseparable models, nonparametric identification, instrumental variables, quantile regression, selection, unobserved heterogeneity, copulas

1 Introduction

The classical linear model, $Y = X\beta + U$, with $\beta \in \mathbb{R}$ and U unobserved, is separable because it assumes that the *ceteris paribus* causal effect of X on Y is β, which is deterministic. The nonparametric model, $Y = m(X) + U$, where m is an unknown function, is also separable because the causal effect of X is $\nabla_x m(X)$, which is deterministic after conditioning on observables. This paper is about the nonseparable model $Y = m(X, U)$. In this model, the causal effect of X on Y is $\nabla_x m(X, U)$, which is still stochastic after conditioning on observables because U is unobserved. This allows the model to capture general unobserved heterogeneity, but also creates an identification problem if X is endogenous, i.e. statistically dependent with U.

Separability is a modeling assumption. For example, if X is an agent's level of schooling and Y is income, then separability implies that all observationally identical agents face the same marginal returns to schooling. If $X \in \{0, 1\}$ is participation in a job training program, separability implies that all observationally identical agents would see the same gain from participating. Assumptions like these are difficult to justify with economic theory. In his Nobel lecture, Heckman (2001) observes that separability is also difficult to justify empirically in many situations. Moreover, because separable models do not allow for unobserved heterogeneity, they are of limited use for evaluating policies from a nuanced standpoint that considers distributional effects (Heckman et al. (1997)).[1]

Nonseparable models typically nest separable models as special cases, which suggests that separability is a source of misspecification. Misspecified models can still be useful if the quantities they identify are reasonably robust to the suspected misspecification. Separable models lack this type of robustness when X is endogenous. Instrumental variable approaches are frequently used to deal with endogeneity in separable models, but if the true model is in fact nonseparable, these methods must be treated with caution. For example, when X is binary, the standard instrumental variable assumptions identify a weighted average of causal effects that is not economically interesting (Heckman and Vytlacil (2005), Heckman et al. (2006)). If the instrument is also binary and the researcher assumes that it has a monotone effect on X, then this

[1]This point has also long been made by proponents of quantile regression such as Koenker (2005).

weighted average becomes interpretable as the local average treatment effect (LATE) of Imbens and Angrist (1994). The definition of the LATE—specifically, the subpopulation for which it is an average—depends on the particular instrument used, so researchers must be careful when extrapolating conclusions based on the LATE to settings without the same instrument. Thus even in basic inference problems, allowing for nonseparability requires additional assumptions about the instrument and, sometimes, qualifies the findings of the analysis.

The internal validity (dependence on the instrument) of the LATE parameter has been criticized as a serious drawback.[2] On the other hand, the assumptions used to identify externally valid features in nonseparable models tend to be strong. Parametric identification strategies, such as in the seminal selection model of Heckman (1979), provide a commonly applied solution. However, while parametric models can be useful in practice, it is desirable to achieve identification independently of the parameterization. An interesting advancement in this direction is made by Florens et al. (2008) who study a nonseparable instrumental variables model in which a single endogenous variable enters the outcome equation in a flexible polynomial structure.

In fully nonparametric models with instrumental variables, the typical identification strategy has been to place nonparametric distributional assumptions on the instrument. Heckman (1990) and Imbens and Newey (2009) consider large support assumptions, which require the instrument to vary over a large set. Such assumptions are not reasonable for many instruments that are used in practice. In particular, they are not satisfied by discrete-valued instruments like intent-to-treat indicators, which arise in randomized experiments with noncompliance. Given the continued growth of randomized experiments in economics, this is an increasingly restrictive shortcoming. A different distributional assumption is the completeness condition studied by Chernozhukov and Hansen (2005) and Chernozhukov et al. (2007). It is not known what the economic content of this assumption is. Finally, there is also a growing body of recent work that uses the time dimension of a panel to nonparametrically identify a nonseparable model, but of course these results presume access to panel data.[3]

In this paper I present conditions under which m is identified in the aforementioned nonseparable model $Y = m(X, U)$, when X and U may be arbitrarily dependent. I

[2]There has been considerable methodological debate about the LATE. See for example the June 2010 issue of *The Journal of Economic Literature* and the Spring 2010 issue of *The Journal of Economic Perspectives*.

[3]Of the papers using panel data, the most closely related to this one is by Evdokimov (2010) because his model is nonparametric and his focus is on identifying the entire outcome equation. Also related is the work of Athey and Imbens (2006), who use repeated cross-sections to identify counterfactual distributions in a nonlinear difference-in-difference framework, and Altonji and Matzkin (2005), who use an exchangeability assumption that is more suited to panels with a group (rather than time) dimension.

consider the special case where Y and X are continuously distributed and m is strictly increasing in U, which is scalar. This model can be interpreted as a nonparametric quantile regression. Identification of nonparametric quantile regressions has been studied recently by Matzkin (2003), Chernozhukov and Hansen (2005) and Chernozhukov et al. (2007). The identification results I provide are externally valid and allow for a general, even binary, instrument. I achieve this by introducing a new assumption that governs the way that the instrument affects the joint distribution of X and U. Formally, I assume that conditioning on the instrument does not affect the copula of these variables. I call this assumption "conditional copula invariance" (CCI). I show that CCI can be satisfied by assuming the existence of a first stage (selection, reduced form) equation that is strictly increasing in a scalar latent variable. I also show that it is satisfied if the instrument acts as an exogenous, rank-preserving treatment on X.

The organization of the paper is as follows. In Section 2, I discuss the model and the assumptions, including CCI. In Section 3, I analyze identification and explain the intuition behind the proof of the main result. In Section 4, I investigate the economic interpretation of CCI and derive low-level sufficient conditions under which it is a valid assumption. I provide examples based on the recent empirical studies of Feinstein and Symons (1999), Hoxby (2000) and Duflo (2001) to demonstrate how CCI can be justified in practice. I also show how my results can be used to derive interpretable, low-level conditions for identification in a nonseparable simultaneous equations model. In Section 5, I construct a nonparametric estimator based on the method of sieves and prove that it is consistent. I provide a Monte Carlo study of this estimator in Section 6. Section 7 concludes.

2 Model

Consider the nonseparable outcome equation

$$Y = m(X, U), \tag{1}$$

where Y is a scalar response variable, $X = (X_1, \ldots, X_{d_x})'$ is a vector of observable explanatory variables and U is a scalar unobservable.[4] I make no assumptions about the marginal dependence between X and U. Let Z denote a scalar or vector-valued instrument that is excluded from (1). For $k = 1, \ldots, d_x$, let $\mathcal{X}_k \subseteq \mathbb{R}, \mathcal{X} \subseteq \mathbb{R}^{d_x}, \mathcal{Z} \subseteq \mathbb{R}^{d_z}, \mathcal{X}_k\mathcal{Z} \subseteq \mathbb{R}^{1+d_z}$ and $\mathcal{X}\mathcal{Z} \subseteq \mathbb{R}^{d_{xz}} = \mathbb{R}^{d_x+d_z}$ denote the supports of $X_k, X, Z, (X_k, Z)$

[4]The results also allow for included explanatory control variables. For exposition, I introduce these in Section 4 and discuss general versions of the results in Appendix A.

and (X, Z), respectively. The function m is unknown and is the parameter of interest. I place the following assumptions on (1).

Assumption M.

M.1 **(Monotonicity)** $m(x, u)$ is continuous in x for every u. For every x, $m(x, u)$ is strictly increasing in u. $Y|(X, Z) = (x, z)$ is continuously distributed for every $(x, z) \in \mathcal{X}\mathcal{Z}$.

M.2 **(Continuity)** $X|Z = z$ is continuously distributed for every $z \in \mathcal{Z}$.

M.3 **(Marginal instrument exogeneity)** $U \perp\!\!\!\perp Z$. (U and Z are independent.)

CCI **(Conditional copula invariance)** Let $C(\cdot, \cdot; z)$ be the copula for $(X, U)|Z = z$ and let $C(\cdot, \cdot)$ be the copula for (X, U).[5] Then $C(\cdot, \cdot; z) = C(\cdot, \cdot)$ for every $z \in \mathcal{Z}$.

There are many important economic applications for the model given by (1) and M.1-M.3. One example is when X is the average class size at a school and Y is a measure of schooling outcomes such as test scores. The object of interest for policy is the causal effect that a class size of $X = x$ has on the distribution of schooling outcomes, $m(x, \cdot)$. The unobservables U will capture, among other things, parental involvement and general student preparedness. The nonseparability of (1) is important because it is likely that the casual effect of class size varies with these unobservables, for example if personal instruction helps bad students more than good students. Since families have some control over the schools that their children attend, sorting on the basis of class size will tend to create a negative dependence between X and U. This means that the causal effect that class size has on outcomes cannot be inferred from observations on X and Y alone because it will be confounded with the effect that the unobservables have on outcomes. An excluded instrument, Z, that satisfies M.3, can potentially identify the causal effect of class size by providing a source of exogenous variation in class size. Two examples of instruments that have been used in the literature are geographic

[5] Sklar's Theorem (originally due to Sklar (1959); see e.g. Nelson (2006)) shows that for any continuously distributed random vector (X, U) with joint distribution F_{XU} there exists a unique probability distribution function C that has domain $[0, 1]^{d_x+1}$ and satisfies

$$F_{XU}(x, u) = C\left(F_{X_1}(x_1), \ldots, F_{X_{d_x}}(x_{d_x}), F_U(u)\right), \text{ for all } x, u.$$

The function C is called a copula, which is Latin for a link or bond. This etymology is natural because C links together the marginal distributions of X and U to create the distribution of (X, U).

Patton (2006) formally extended Sklar's Theorem to a conditional version which concludes similarly that there exists a unique $C(\cdot, \cdot; z)$ such that

$$F_{XU|Z}(x, u \mid z) = C\left(F_{X_1|Z}(x_1 \mid z), \ldots, F_{X_{d_x}|Z}(x_{d_x} \mid z), F_{U|Z}(u \mid z); z\right), \text{ for all } x, u,$$

where Z is any random element and the conditioning is on the event $[Z = z]$ for $z \in \mathcal{Z}$.

indicators (for large regions) and exogenous fluctuations in student enrollment due to birth patterns. I discuss these instruments more in Section 4.2, with a focus on the interpretation of the additional assumption CCI.

To identify $m(x, \cdot)$ at some $x \in \mathcal{X}$ of interest, I also require the following mild rank condition. The content and motivation of this assumption is discussed in Section 3, where it is reformulated as "local dependence." Here and throughout the paper, terms like $F_{X|Z}(\cdot \mid z)$ denote conditional distribution functions for the event $[Z = z]$ with $z \in \mathcal{Z}$.

Assumption R. (Rank condition) There is an $\epsilon > 0$ such that for a.e. $x' \in B_\epsilon(x)$, $\mathbb{P}\left\{z \in \mathcal{Z} : F_{X_k|Z}(x'_k \mid z) \neq F_{X_k}(x'_k)\right\} > 0$ for $k = 1, \ldots, d_x$.

Matzkin (2003) showed that m is inherently unidentified without an additional normalization to fix the units of measurement. All of the normalizations she considers are applicable to this paper. To motivate the model, I consider the following normalization as the leading case.[6]

N.1 (Quantile regression normalization) $U \sim \text{Unif}[0, 1]$. (U is distributed uniformly on $[0, 1]$.)

As a function of u, $m(x, \cdot)$ describes the distribution of potential outcomes of Y when X is exogenously set to x. The u^{th} quantile of this counterfactual distribution is called the quantile treatment response (QTR) to x by Chernozhukov and Hansen (2005). Under M.1 and N.1, the u^{th} QTR to x is $m(x, u)$. The quantile treatment effect (QTE) of x' relative to x is the impact on the u^{th} quantile of potential outcomes due to an exogenous shift from x to x' and is given by $m(x', u) - m(x, u)$.[7] When $U \not\perp\!\!\!\perp X$, the QTR to x is generally different than the conditional quantile of Y, i.e. $Q_{Y|X}(u \,|\, x) = m\left(x, Q_{U|X}(u \mid x)\right) \neq m(x, u)$, where the first equality follows from M.1.[8] Similarly, the QTE of x' relative to x is not equal to the difference of the conditional quantiles of Y at $X = x'$ and $X = x$.

The model can be written in the usual quantile regression notation as

$$Y = m_s(X) + U_s, \tag{2}$$

[6] All identification results are given in forms invariant to the normalization used. I discuss other normalizations in Appendix B.

[7] The original formulation of the QTE for binary treatments is due to Lehmann (1974) and Doksum (1974).

[8] For any scalar random variable Y and random element X, $Q_Y(s) = \inf\{y : F_Y(y) \geq s\}$ and $Q_{Y|X}(s|x) = \inf\left\{y : F_{Y|X}(y \mid x) \geq s\right\}$.

where $m_s(X) = m(X, s)$ and $U_s = m(X, U) - m(X, s)$. Unlike exogenous nonparametric quantile regression (e.g., Chaudhuri (1991)), the unobservable U_s need not satisfy the condition $Q_{U_s|X}(s \,|\, X) = 0$. The notation in (2) is natural for standard quantile regression analysis, which fixes a given quantile of interest, s. Chernozhukov and Hansen (2005) and Chernozhukov et al. (2007) show that global and local identification of m_s can be achieved through nonlinear versions of the well-known completeness condition of Newey and Powell (2003), i.e. $\mathbb{E}\left(\Delta(X) \,|\, Z\right) = 0$ if and only if $\Delta(X) = 0$ almost surely. The nonlinear completeness conditions place restrictions on the unobservable distribution of (X, U, Z), whereas the Newey and Powell (2003) restriction only concerns the distribution of (X, Z). The nonlinear conditions also depend on the true m_s that generated the data. As a result, it is not clear what the economic interpretation of nonlinear completeness is.[9] CCI serves as an alternative identifying assumption. As will become clear, it places restrictions across different quantile equations in (2) simultaneously. This makes the traditional notation clumsy, which is why I focus on (1) rather than (2).

Hoderlein and Mammen (2009) argue that M.1 is restrictive for structural applications since it allows for only a single, scalar unobservable. Matzkin (2003) shows that there are some structural models of economic interest where the outcome equation has only a single unobservable, but in general the point of Hoderlein and Mammen (2009) is well-taken. However, Chesher (2007) observes that when U is not scalar one must either be content with identifying average quantities or be willing to model and estimate additional observable equations.[10] This is problematic, because it is not clear that these average quantities have useful structural interpretations, which is the impetus for increasing the dimension of U to begin with. The dimension of the unobservable in a nonseparable model thus generates an interesting friction that is still being investigated in econometric research.

Under N.1, Assumption M.1 has the effect of aggregating all unobserved factors into a single random variable that captures the quantile ranking among observationally identical units. It also imposes rank invariance, a concept introduced by Doksum (1974), which assumes that the realizations of U do not change if X is counterfactually manipulated.[11] If the analyst's goal is to identify distributional effects, rank invariance

[9] The main contribution of Chernozhukov and Hansen (2005) is for the case when X is discrete. Their results for this case do not depend on a completeness condition. In particular, the authors provide natural sufficient conditions for global identification in the important case when both X and Z are binary.

[10] Imbens and Newey (2009) identify averages for a vector U. Chesher (2003) takes the other approach of modeling one observable equation for each dimension of the unobservable.

[11] It turns out that rank invariance is also important for interpreting Assumption CCI, so I discuss it more in Section 4.

seems difficult to relax in any meaningful way without more specific behavioral assumptions, either in this model or in similar models in the literature. This is to be expected since, by definition, the data is not informative about counterfactual outcomes. It is also really the same point made by Chesher (2007) regarding non-scalar U, but with a different interpretation. Heckman et al. (1997) provide an insightful account of several issues related to rank invariance, including some economically motivated assumptions which may be helpful in relaxing it. See also the discussions in Chernozhukov and Hansen (2005) and Athey and Imbens (2006).

The key part of Assumption M is conditional copula invariance (CCI) which requires that the instrument does not affect the form of the copula of $(X,U)|Z$.[12] The copula of a random vector characterizes the dependence structure between its component variables. CCI can therefore be interpreted as saying that the instrument does not affect the underlying cause of the endogeneity of X. Nonseparable instrumental variable models tend to require strong identifying assumptions (or yield weak conclusions) precisely because M.3 is not enough to make an excluded instrument fully exogenous. The reason is that although a change in Z does not affect the marginal distribution of U, it could still change the joint distribution of $(X,U)|Z$ by affecting the copula, which is also unobserved. CCI is the assumption that is needed to solve this problem.

To clarify this point, assume for the sake of exposition that X is a scalar and Z is continuously distributed with $F_{XU|Z}$ differentiable in Z. The conditional form of Sklar's Theorem (footnote 5) shows that the total effect of Z on $F_{XU|Z}$ can be uniquely decomposed into three distinct components,

$$\begin{aligned}\nabla_z F_{XU|Z}(x,u\mid z) &= \nabla_z C\left(F_{X|Z}(x\mid z), F_{U|Z}(u\mid z); z\right)\\ &= \quad C_1\left(F_{X|Z}(x\mid z), F_{U|Z}(u\mid z); z\right)\nabla_z F_{X|Z}(x\mid z)\\ &\quad + C_2\left(F_{X|Z}(x\mid z), F_{U|Z}(u\mid z); z\right)\nabla_z F_{U|Z}(u\mid z)\\ &\quad + C_z\left(F_{X|Z}(x\mid z), F_{U|Z}(u\mid z); z\right), \qquad (3)\end{aligned}$$

where $C_i, i=1,2,z$ are the derivatives of the copula of $(X,U)|Z$ with respect to its first, second and third arguments. The first and second terms of (3) are the effects that Z has on $F_{XU|Z}$ through its impact on the marginal distributions of $X|Z$ and $U|Z$. Notice that under M.3 and N.1, $F_{U|Z}(u\mid z)=u$. This causes the second term to

[12] Bond and Shaw (2006) used a slightly stronger assumption to achieve partial identification in a competing risks model. Abbring and van den Berg (2005) applied the results of Bond and Shaw (2006) to a duration model with dynamic treatment effects. Other work on competing risks, such as Heckman and Honoré (1989) and Abbring and van den Berg (2003), has implicitly used CCI. To my knowledge, this paper is the first to investigate the implications of CCI in a nonseparable model.

vanish. The third term is the effect that Z has on the marginal-invariant dependence structure between X and U, or, put differently, on the form of the copula of $(X, U)|Z$. CCI causes the third term to vanish as well. Thus, the combined effect of M.3 and CCI is to require that the instrument not affect the unobservable components of $F_{XU|Z}$. Specifically, these assumptions with N.1 imply that

$$\nabla_z F_{XU|Z}(x, u \mid z) = C_1\left(F_{X|Z}(x \mid z), u\right) \nabla_z F_{X|Z}(x \mid z).$$

The only unobservable part of this expression is the copula, which does not depend on z by CCI. It is important to note that these assumptions still allow the observable component of the joint distribution of $F_{XU|Z}$ to vary with z. Namely, $F_{X|Z}$ need not equal F_X, and in fact Assumption R requires that they are not equal.[13]

The following equivalence is referred to frequently throughout the paper and is central to the intuition.

Proposition 1. *Define $R_k = F_{X_k|Z}(X_k \mid Z)$. Then given Assumptions M.2-M.3, CCI holds if and only if for every k, $U \perp\!\!\!\perp Z|R_k$.*[14]

To understand Proposition 1 it is useful to appeal to the informational interpretation of statistical independence (e.g. Section 3.8 of Rényi (1970)). Suppose $d_x = 1$. The random variable R, or the "conditional rank" of X, provides information on X because under M.2 the event $[R = r]$ is the same as the event $[X = Q_{X|Z}(r \mid Z)]$. Since the model allows for $U \not\perp\!\!\!\perp X$, R therefore also provides information on U. One might think that Z provides additional information about U that is not encoded in R through the same logic; Z carries information on X which is dependent on U. Proposition 1 shows that this possibility is assumed away by CCI. In other words, the instrument contains no information about U that is not already contained in the conditional rank of X. CCI is also sufficient, though not necessary, for the conditional rank to be a control function for X.[15]

Proposition 2. *Given Assumptions M.2-M.3, CCI implies that for every k, $U \perp\!\!\!\perp X_k|R_k$. The converse is not true.*

[13] Applications of copula theory are almost always parametric, but the decomposition in (3) demonstrates the utility of the theory as a purely analytic tool. CCI can be restated without reference to copulas as

$$\nabla_z F_{XU|Z}(x, u \mid z) = \quad F_{U|XZ}(u \mid x, z) \nabla_z F_{X|Z}(x \mid z) + F_{X|UZ}(x \mid u, z) \nabla_z F_{U|Z}(u \mid z),$$

but this is unwieldy and non-intuitive. The analogous statement for discrete Z would be even worse.

[14] All proofs are contained in the appendices.

[15] See Section 2.2 of Blundell and Powell (2003) for a survey of control function methods.

For intuition, continue to assume that $d_x = 1$. The combined implication of Propositions 1 and 2 is that CCI requires R to be a control function not just for X, but also for Z. This may seem an unusual statement since M.3 suggests that Z should not need to be controlled for. This apparent conflict underscores a crucial difference in the workings of instrumental variable methods between separable and nonseparable models. In both separable and nonseparable models, Z affects the outcome, Y, through its effect on X, given an appropriate rank condition such as Assumption R. Also in both types of models, Z can affect not just the marginal distribution of X but also the marginal-invariant component of the joint distribution of X and U, i.e. the copula of $(X, U)|Z$. In a separable model, the impact that Z has on the copula of $(X, U)|Z$ is irrelevant for recovering the causal effect of X on Y. This is simply because a separable model assumes that the causal effect of X on Y does not depend on U. Nonseparable models relax this assumption, which then requires consideration of how the instrument affects the copula. Propositions 1 and 2 show that CCI allows one to use R to control for both X and the additional copula effect.[16]

In Section 4 I use Proposition 1 to derive low-level sufficient conditions for CCI in terms of the relationship between X, Z and U. It turns out that these conditions nest some commonly used first stage equations and counterfactual conditions. In that section I also provide economic examples which explore when CCI can be expected to hold and when it may be a bad assumption. These examples suggest that in many contexts CCI functions as a relatively weak addition to M.3.

3 Identification

In this section I show that Assumptions M and R provide for identification of $\bar{m}(x, s) = m(x, Q_U(s))$ for $s \in [0, 1]$. Under the quantile regression normalization N.1, this shows that $m(x, u)$ is identified for $u \in [0, 1]$. In Appendix B I consider alternative normalizations that provide for identification of $m(x, \cdot)$ from $\bar{m}(x, \cdot)$.

Consider the following proposition, which does not rely on the existence of an excluded instrument. It shows that identification of $\bar{m}$ is equivalent to identification of the copula, C.[17]

[16] When $d_x > 1$, the preceding control function intuition becomes more nuanced because there are d_x control functions $R_1, \ldots, R_{d_x}$—one for each X_k. In general, this does not allow one to interpret $R = (R_1, \ldots, R_{d_x})$ as a control function for the entire vector X, due to dependence among the components of R.

[17] Here and throughout I will not distinguish between the identification of $C_{\mathbf{X}}$ and C. They are equivalent, since $\mathbf{F}_X$ is observable. See the proof of Proposition 1 for a formal justification.

Proposition 3. *Under Assumptions M.1 and M.2, identification of $\bar{m}(x,s)$ is equivalent to identification of $C_{\mathbf{X}}(\mathbf{F}_X(x),s)$, where $\mathbf{F}_X(x) = (F_{X_1}(x_1),\ldots,F_{X_{d_x}}(x_{d_x}))'$ is the d_x-dimensional vector of marginal distributions and $C_{\mathbf{X}}(r,s) = \nabla^{d_x}_{r_1,\ldots,r_{d_x}} C(r,s)$ for $(r,s) \in (0,1)^{d_x+1}$ is the d_x^{th}-order partial derivative of C with respect to each element of r. In particular,*

$$C_{\mathbf{X}}\left(\mathbf{F}_X(x),s\right) = F_{Y|X}\left(\bar{m}(x,s) \mid x\right)/\sigma(x) \qquad \textit{and}$$
$$\bar{m}(x,s) = Q_{Y|X}\left(C_{\mathbf{X}}\left(\mathbf{F}_X(x),s\right)\sigma(x) \mid x\right)$$

for all $x \in \mathcal{X}$ and $s \in (0,1)$, where

$$\sigma(x) = \frac{\prod_{k=1}^{d_x} f_{X_k}(x_k)}{f_X(x)}$$

is an observable scaling factor composed of the marginal and joint densities of X. Note that $\sigma(x) = 1$ if $d_x = 1$ or if the components of X are jointly independent.

This result extends an argument due to Matzkin (2003), who considered identification in the special case when $U \perp\!\!\!\perp X$. Since $C_{\mathbf{X}}\left(\mathbf{F}_X(x), F_U(u)\right)\sigma(x) = F_{U|X}(u \mid x)$,[18] Proposition 3 shows that $\bar{m}(x,s) = Q_{Y|X}\left(F_{U|X}(Q_U(s) \mid x) \,\middle|\, x\right) = Q_{Y|X}(s \mid x)$ is constructively identified in this case. From there, only a normalization is needed to identify $m(x,u)$. Matzkin (2003) also considers the weaker independence condition $U \perp\!\!\!\perp X|W$ and $U \perp\!\!\!\perp W$, where W is a set of observable control variables. In Appendix A I show that Proposition 3 adapts to cover this case as well. The weaker condition amounts to the assumption that W comprises a set of exogenous controls rich enough to overcome the endogeneity between X and U, a statement typically called unconfoundedness or selection on observables.[19] In contrast, this paper allows for selection on unobservables, $X \not\perp\!\!\!\perp U|W$, which is desirable in cases where U is thought to affect the determination of X, often through the decision process of an economic agent. Assumptions M.1 and M.2 are not adequate to identify $\bar{m}$ in this case.

The benefit of an excluded instrument that satisfies M.3 and CCI is derived from the following proposition, which uses an argument similar to that in Proposition 3.

[18] This is established in Proposition C.1.

[19] In fact, if the causal effect of W is not of interest then these controls do not need to be exogenous to identify the causal effect of X. I discuss this in Appendix A.

Proposition 4. *Under Assumptions M.1-M.3 and CCI,*

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right) = F_{Y|XZ}\left(\bar{m}(x,s) \mid x, z\right)/\sigma(x|z) \text{ and}$$
$$\bar{m}(x,s) = Q_{Y|XZ}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\sigma(x|z) \,\middle|\, x, z\right)$$

for all $(x,z) \in \mathcal{XZ}, s \in (0,1)$*, where* $\mathbf{F}_{X|Z}(x \mid z) = (F_{X_1|Z}(x_1 \mid z), \ldots, F_{X_{d_x}|Z}(x_{d_x} \mid z))'$ *and* $\sigma(x|z) = \left(\prod_{k=1}^{d_x} f_{X_k|Z}(x_k \mid z)\right)/f_{X|Z}(x \mid z)$ *is a scaling factor composed of the marginal and joint densities of* X*, conditional on* Z*.*

By equating $\bar{m}(x,s)$ in Propositions 3 and 4, it follows that

$$Q_{Y|X}\left(C_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\sigma(x) \mid x\right) = Q_{Y|XZ}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\sigma(x|z) \,\middle|\, x, z\right) \tag{4}$$

for all $(x,z) \in \mathcal{XZ}$ and all $s \in (0,1)$. Notice that for a candidate parameter value, C', and knowledge of the joint distribution of the observable random vector, (Y,X,Z), (4) can be computed for any s. This equation is therefore an observable implication of Assumption M. Any copula that satisfies the assumptions of the model must satisfy (4). An alternative to equating $\bar{m}$ in Propositions 3 and 4 is to equate the copula, properly adjusted to be evaluated at the same point in both equations. This leads to the equivalent observable implication

$$\bar{m}(x,s) = Q_{Y|X}\left(F_{Y|XZ}\left(\bar{m}\left[q(x|z), s\right] \mid q(x|z), z\right)\frac{\sigma(x)}{\sigma\left[q(x|z)|z\right]} \,\middle|\, x\right), \tag{5}$$

where $q_k(x_k|z) = Q_{X_k|Z}\left(F_{X_k}(x_k) \mid z\right)$ and $q(x|z) = (q_1(x_1|z), \ldots, q_{d_x}(x_{d_x}|z))'$. For the purpose of identification it is more intuitive to analyze (4). However (5) has advantages for estimation, so I return to it in Section 5.

Let $\mathcal{C}$ be the set of all (d_x+1)-dimensional copulas and define the mapping $\Gamma : \mathcal{C} \times \mathcal{XZ} \to \mathbb{R}$ by

$$\begin{aligned}\Gamma(C', x, z) = \big\| & Q_{Y|X}\left(C'_{\mathbf{X}}(\mathbf{F}_X(x), \cdot)\sigma(x) \,\middle|\, x\right) \\ & - Q_{Y|XZ}\left(C'_{\mathbf{X}}(\mathbf{F}_{X|Z}(x \mid z), \cdot)\sigma(x|z) \,\middle|\, x, z\right)\big\|_U^2,\end{aligned}$$

where $\|\cdot\|_U$ is the L_2-norm with respect to Lebesgue measure on the unit interval $(0,1)$. Equality (4) implies that $\Gamma(C, x, z) = 0$ for every $(x,z) \in \mathcal{XZ}$, where C is the true copula in the data generating process. This suggests considering the criterion function

$$T : \mathcal{C} \to \mathbb{R} : T(C') = \mathbb{E}\,\Gamma(C', X, Z). \tag{6}$$

If $T(C') = 0$ only for $C' = C$ then T identifies the model. Since $\Gamma \geq 0$, $T(C') = 0$ if and only if $\Gamma(C', x, z) = 0$ a.s.-F_{XZ}. Hence if $T(C') = 0$ then

$$Q_{Y|X}\left(C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\sigma(x) \,\middle|\, x\right) = Q_{Y|XZ}\left(C'_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\sigma(x|z) \,\middle|\, x, z\right) \tag{7}$$

for every $s \in (0,1)$ and a.e. $(x,z) \in \mathcal{X}\mathcal{Z}$.[20]

Propositions 3 and 4 established that $Q_{Y|X}$ and $Q_{Y|XZ}$ can be expressed in terms of C, the true copula that is assumed to generate the data. Letting $t = C_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\sigma(x)$ in Proposition 3 and $t = C_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\sigma(x|z)$ in Proposition 4, it follows that

$$Q_{Y|X}(t \mid x) = \bar{m}\left(x, C_{\mathbf{X}}^{-1}\left[\mathbf{F}_X(x), t/\sigma(x)\right]\right) \quad \text{and} \tag{8}$$

$$Q_{Y|X,Z}(t \mid x, z) = \bar{m}\left(x, C_{\mathbf{X}}^{-1}\left[\mathbf{F}_{X|Z}(x \mid z), t/\sigma(x|z)\right]\right), \tag{9}$$

where $C_{\mathbf{X}}^{-1}$ is the inverse of $C_{\mathbf{X}}$ in its $(d_x+1)^{\text{th}}$ argument.[21] Applying relationships (8)-(9) to (7) yields

$$\begin{aligned}&\bar{m}\left(x, C_{\mathbf{X}}^{-1}\left[\mathbf{F}_X(x), C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\right]\right)\\ &\quad = \bar{m}\left(x, C_{\mathbf{X}}^{-1}\left[\mathbf{F}_{X|Z}(x \mid z), C'_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\right]\right).\end{aligned} \tag{10}$$

Since $\bar{m}(x, \cdot)$ is strictly increasing by M.1, (10) can hold if and only if

$$C_{\mathbf{X}}^{-1}\left(\mathbf{F}_X(x), C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\right) = C_{\mathbf{X}}^{-1}\left(\mathbf{F}_{X|Z}(x \mid z), C'_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\right). \tag{11}$$

In summary, under Assumption M, $T(C') = 0$ if and only if (11) holds for every $s \in (0,1)$ and a.e. $(x,z) \in \mathcal{X}\mathcal{Z}$.

For (11) to be informative, it must be the case that $\mathbf{F}_X \neq \mathbf{F}_{X|Z}$. Otherwise, if $\mathbf{F}_X(x) = \mathbf{F}_{X|Z}(x \mid z)$ for some $x \in \mathcal{X}$ and all $z \in \mathcal{Z}$, (11) is a tautology when evaluated at x. In other words, a rank condition is needed to ensure the relevance of the instrument. Rank conditions require it not to be the case that $X \perp\!\!\!\perp Z$, with more primitive conditions being particular to the model at hand. While $X \perp\!\!\!\perp Z$ has a precise global meaning, the meaning of its negation is necessarily local. This is the reason for

[20] There is a substantial amount of flexibility in specifying the objective function (6). For example, a strictly positive weighting function $\omega(x,z)$ could be included in the moment so that $T(C') = \mathbb{E}\,\omega(X,Z)\Gamma(C',X,Z)$. Also, the definition of Γ could be any functional for which $\Gamma(C',x,z) = 0$ implies (7). For identification, the specific choice is irrelevant. It is important for estimation, but investigation of an optimal choice of Γ and ω is beyond the scope of this paper. Manski (1983) and Komunjer and Santos (2010) also examine estimation problems with this feature.

[21] That is, for all $(r,t) \in (0,1)^{d_x+1}$, $C_{\mathbf{X}}(r, C_{\mathbf{X}}^{-1}(r,t)) = t$. Assumption M.1 provides that $U|X = x$ is continuously distributed for any x and hence that $C_{\mathbf{X}}^{-1}(r, \cdot)$ exists for any $r \in (0,1)^{d_x}$.

introducing the following concept.

Definition 1 (**Local independence**)**.** Consider any k, any $x_k \in \mathcal{X}_k$ and the set

$$\mathcal{Z}_k(x_k) = \{z \in \mathcal{Z} : F_{X_k}(x_k) \neq F_{X_k|Z}(x_k \mid z)\}.$$

X_k is *locally dependent* on Z at x_k if $\mathbb{P}\,\mathcal{Z}_k(x_k) > 0$, expressed as "$X_k \not\perp\!\!\!\perp Z$ at x_k." Otherwise X_k is *locally independent* of Z at x_k, expressed as "$X_k \perp\!\!\!\perp Z$ at x_k."

The d_x-dimensional vector X is defined to be locally dependent on Z at x, i.e. "$X \not\perp\!\!\!\perp Z$ at x" if $X_k \not\perp\!\!\!\perp Z$ at x_k for every k.

Given M.2, the rank condition that was briefly mentioned in Section 2 can be conveniently reformulated in terms of local dependence.

Assumption R. (Local rank condition) $X \not\perp\!\!\!\perp Z$ at x.

Assumption R is weak and, as I will argue, minimal. It assumes nothing about the magnitude of the dependence between X and Z, only that there is some dependence at x. Suppose that $d_x = 1$ and that $X = \pi Z + V$, where $V \perp\!\!\!\perp Z$ is continuously distributed. A necessary condition for Assumption R is that $\pi \neq 0$. Imbens and Newey (2009), who consider a control function approach for a similar nonseparable model, also require $\pi \neq 0$, but in addition they require the large support assumption that $\mathcal{Z} = \mathbb{R}$. This assumption rules out discrete instruments and its credibility can be strained even for continuous instruments.[22] In contrast, Assumption R allows for a binary instrument. It also does not require $d_z \geq d_x$, the familiar order condition from the linear model.

To explain the sense in which Assumption R is minimal, consider the famous study of Angrist and Krueger (1991) (AK) about the effect of schooling on lifetime income and the recent re-analyses of Chesher (2005, 2007). AK use an individual's quarter of birth as an instrument for their schooling, with the argument that date of birth is exogenous to lifetime income, but could affect schooling because of compulsory education laws. Using a large sample, they find a statistically significant correlation between quarter of birth and schooling.[23] Chesher employs a nonparametric, nonseparable local approach to show that the analysis in AK depends strongly on the assumed linear model. He shows that linearity allows point estimates that may be roughly valid for agents of very low schooling to be extrapolated out to those with higher schooling, for whom

[22] Imbens and Newey (2009) also provide sharp set identification results for their model when the support of Z is not large enough to obtain point identification.

[23] AK has been a motivation for the important literature on weak instruments. This issue has no relevance to the current discussion of identification, although it bears mentioning that the rank condition in this paper does not suggest an obvious metric of instrument weakness.

compulsory education laws are not binding and the quarter of birth instrument is not relevant.

Assumption R requires the analysis to focus only on the schooling levels for which the quarter of birth instrument has an effect. Thus Assumption R can be seen as an explicit "no extrapolation" condition. This is a reasonable demand to make of an instrument if it is to be the source of identification in a nonseparable, nonparametric model. If the condition fails for the desired instrument at the desired choice of $x \in \mathcal{X}$ then the analyst can either a) change or augment the instrument, b) limit the analysis to the $x \in \mathcal{X}$ on which Assumption R does hold, or c) impose additional structure on the model in order to make extrapolation possible.[24] While none of these alternatives are desirable, there is a sense in which they partition the logical options in empirical analysis: get different data, make more limited statements, or impose more (or different) assumptions. Assumption R can therefore be called a minimal rank condition for a nonparametric, nonseparable model.

The following theorem is the main result of the paper. I emphasize that the only substantive assumptions on Z required for identification in Theorem 1 are M.3 (marginal exogeneity), CCI and the rank condition, Assumption R. In particular, Z can be a binary random variable.

Theorem 1. *Under Assumptions M and R, $T(C') = 0$ if and only if $C'_{\mathbf{X}}(\mathbf{F}_X(x), s) = C_{\mathbf{X}}(\mathbf{F}_X(x), s)$ for every $s \in [0,1]$. Hence by Proposition 3, $\bar{m}(x,s)$ is identified. If N.1 is maintained, then this shows that $m(x,u)$ is identified.*

Let $\mathcal{X}^D = \{x \in \mathcal{X} : X \not\perp\!\!\!\perp Z \text{ at } x\}$ be the (measurable) subset of the support for which Assumption R holds and assume that $\mathbb{P}\,\mathcal{X}^D > 0$.[25] Assumption R will always hold almost surely for the random variable $X^D = X | X \in \mathcal{X}^D$. The proof of identification makes use of the implication of Assumption M that (11) holds almost everywhere. It also requires Assumption R to hold almost everywhere. To justify my statements of Assumption R as a local condition and Theorem 1 as a local result, it is therefore important to show that if Assumption M holds with X then it also holds when X is replaced by X^D. Assumptions M.1-M.3 are clearly inherited by X^D. The following proposition shows that CCI is as well.

[24] Suppose that C (equivalently, m) is assumed to be a member of a parametric family $\{C(\cdot,\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{d_\theta}\}$ and that the data is generated by some $\theta_0 \in \Theta$. As detailed in the upcoming discussion, Assumption R will generally hold on some non-negligible subset of $\mathcal{X}$. Hence, Theorem 1 provides identification of θ_0, as long as the parametric family is well-behaved on the subset for which Assumption R holds. Given the parametric assumption, this allows for extrapolation over all of $\mathcal{X}$.

[25] This is an extremely weak assumption. When $d_x = 1$, M.2 guarantees that $\mathbb{P}\,\mathcal{X}^D > 0$ as long as the instrument is not completely irrelevant, i.e. $X \perp\!\!\!\perp Z$.

Proposition 5. *If CCI holds for (X, U, Z) then it also holds for (X^D, U, Z).*

To avoid continually making the distinction between X and X^D, I will assume for the rest of this section that $\mathcal{X} = \mathcal{X}^D$, i.e. $\mathbb{P}\,\mathcal{X}^D = 1$.

In the following, I give the proof of Theorem 1 for a special case and explain the intuition for the general case. Noting that (11) does not depend on z, define $I(x, s) = (11)$, where the letter "I" is meant to suggest "identity." The strategy of the proof of Theorem 1 is to show that the only way that (11) can hold for a.e. $(x, z) \in \mathcal{X}\mathcal{Z}$ and every $s \in (0, 1)$ is if $I(x, s) = s$ for any x such that $X \not\perp\!\!\!\perp Z$ at x. If this is the case then (11) expresses the identity relationship $C_{\mathbf{X}}^{-1}(\mathbf{F}_X(x), \cdot) \circ C'_{\mathbf{X}}(\mathbf{F}_X(x), \cdot) = id$, i.e. $C_{\mathbf{X}}(\mathbf{F}_X(x), \cdot) = C'_{\mathbf{X}}(\mathbf{F}_X(x), \cdot)$. This is identification of C and hence identification of $\bar{m}$, by Proposition 3. Showing $I(x, s) = s$ proceeds in two steps.

The first step is to establish that $I(x, s) = I(s)$ is not a function of x for $x \in \mathcal{X}^D$. This is somewhat delicate when Z is potentially discrete, but when Z is continuously distributed with connected support, $F_{X|Z}$ is differentiable in z and X is a scalar ($d_x = 1$), it follows by differentiating $I(x, s)$ with respect to z and with respect to x,

$$\begin{aligned} 0 = \nabla_z I(x, s) &= \left(\nabla_1 C_1^{-1}\right) \nabla_z F_{X|Z} + \left(\nabla_2 C_1^{-1}\right) \left(\nabla_1 C_1'\right) \nabla_z F_{X|Z}, \\ \nabla_x I(x, s) &= \left(\nabla_1 C_1^{-1}\right) f_{X|Z} + \left(\nabla_2 C_1^{-1}\right) \left(\nabla_1 C_1'\right) f_{X|Z}, \end{aligned}$$

where $\nabla_i C_1^{-1}, i = 1, 2$ are the derivatives of C_1^{-1} with respect to its first and second components and all points of evaluation are suppressed for readability. If $X \not\perp\!\!\!\perp Z$ at x, then given the smoothness assumptions made on the instrument, $\nabla_z F_{X|Z}(x \mid z)$ is non-zero for some $z \in \mathcal{Z}$. Hence $\left(\nabla_1 C_1^{-1}\right) + \left(\nabla_2 C_1^{-1}\right)\left(\nabla_1 C_1'\right) = 0$ and $\nabla_x I(x, s) = 0$ as well.

The second step is to show that if $I(x, s) = I(s)$ is only a function of s, then because C and C' are proper copulas, it must be the case that $I(s) = s$. This follows by rewriting (11) as $C_1(F_X(x), I(s)) = C_1'(F_X(x), s)$, weighting each side by the marginal density of X, and then integrating over $\mathcal{X}^D$. Since C, C' are proper copulas—specifically, they have unit-uniform marginal distributions—and because $\mathbb{P}\,\mathcal{X}^D = 1$, as assumed for simplicity, this yields $I(s) = s$.

An intuitive requirement for identification of levels in a nonparametric model is that the assumptions are sufficient to provide a comparison of the causal effects of x_a and x_b on Y for any $x_a, x_b \in \mathcal{X}$. This can be demonstrated here through a "chaining" argument, which gives intuition for why the result holds for general Z. For simplicity, assume that X is scalar and that $\mathcal{X}\mathcal{Z} = \mathcal{X} \times \mathcal{Z}$. The argument makes use of Proposition 1, which showed that $U \perp\!\!\!\perp Z | R$. This can be interpreted as saying that changes in Z are causal after conditioning on R. Assumptions M.2 and R imply that conditional

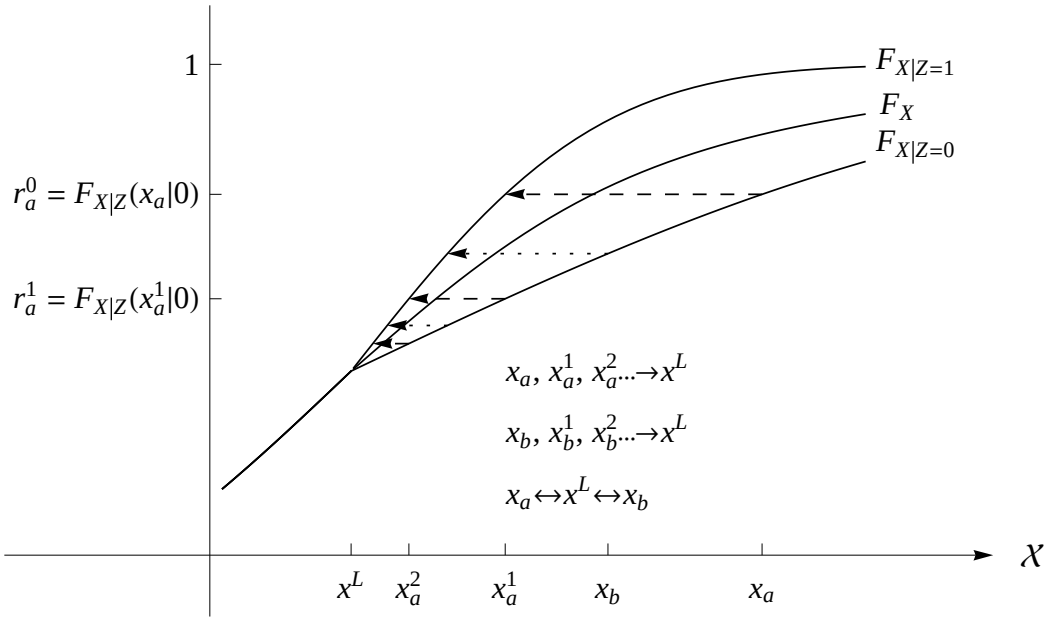

Figure 1: **Chaining together multiple causal comparisons when Z is binary.** The dashed arrows represent causal shifts starting from x_a while the dotted arrows represent causal shifts starting from x_b. The rank condition is only satisfied here for $x \geq x^L$, so this instrument cannot identify $\bar{m}$ for $x < x^L$ without further extrapolating assumptions.

on $R = r \in (0, 1)$, a causal shift of Z from z to z' corresponds to a unique causal shift between two distinct points in $\mathcal{X}$. Thus, given the exclusion of Z from (1), any difference in the distributions of $Y|R = r, Z = z$ and $Y|R = r, Z = z'$ is entirely attributable to the associated shift in X. When Z is binary, only one such causal shift is possible for any given r. However, M.2 allows multiple causal comparisons to be linked together so that any two arbitrary points can be compared by repeated shifts of Z.

For example, to compare the causal effects of x_a and x_b when Z is binary and X is scalar, first find the r_a^0 such that $x_a = Q_{X|Z}\left(r_a^0 \,\middle|\, 0\right)$. Such an r_a^0 exists by M.2. Now shift $Z = 0$ to $Z = 1$ while fixing r_a^0. This represents a causal shift from x_a to $x_a^1 = Q_{X|Z}\left(r_a^0 \,\middle|\, 1\right)$. If Assumption R holds at x_a then $x_a^1 \neq x_a$, so this causal comparison is not trivial. Next, find the r_a^1 such that $x_a^1 = Q_{X|Z}\left(r_a^1 \,\middle|\, 0\right)$. Repeat the process by finding an $x_a^2 = Q_{X|Z}\left(r_a^1 \,\middle|\, 1\right)$, which gives a causal shift from x_a^1 to x_a^2. Notice that now x_a has been causally compared with x_a^2 through their mutual comparisons with x_a^1. The proof of Theorem 1 shows that this sequence x_a, x_a^1, x_a^2 can be continued indefinitely and that the continuity of X allows x_a to be causally compared with a unique limiting point, x^L. Moreover, an analogous sequence started at x_b has the same limit. Thus x_a and x_b can be causally compared through their

mutual comparisons with x^L. This is the intuition behind Theorem 1. Figure 1 depicts this argument graphically.

4 Conditional Copula Invariance

4.1 Sufficient conditions

CCI is a nonparametric assumption about the relationship between X, U and Z. In this section I investigate the economic content of this assumption and justify it in two different, but related ways. The first way provides sufficient conditions on a first stage (selection, reduced form) equation that describes how Z affects X. This approach is natural if a structural model of the instrument's effect can be formulated. The second way phrases sufficient conditions using the language and notation of causal statistics, which has also found currency in the recent microeconometric literature. The counterfactual notation used in this literature lends itself to more precisely stated assumptions, but must be justified through intuitive arguments. Depending on the application, this may be more or less natural than constructing a first stage equation, as I demonstrate through the examples in Section 4.2. The economic interpretation under either set of sufficient conditions is closely related, even though the conditions themselves are not precisely equivalent.

Throughout this section I explicitly condition on covariates W that may be included in outcome equation (1) and may or may not be independent of U. Covariates have an important role in the interpretation of M.3 and CCI. Typically, they will make these assumptions more reasonable. Assumption M needs to be modified as follows. For these assumptions, $Y = m(X, W, U)$, and $\mathcal{WZ}, \mathcal{XWZ}$ respectively denote the supports of (W, Z) and (X, W, Z).

Assumption M′.

M.1′ **(Monotonicity)** $m(x, w, u)$ is continuous in x for every w, u. For every x, w, $m(x, w, u)$ is strictly increasing in u. $Y|(X, W, Z) = (x, w, z)$ is continuously distributed for every $(x, w, z) \in \mathcal{XWZ}$.

M.2′ **(Continuity)** $X|(W, Z) = (w, z)$ is continuously distributed for every $(w, z) \in \mathcal{WZ}$.

M.3′ **(Marginal instrument exogeneity, conditional on covariates)** $U \perp\!\!\!\perp Z|W$.

CCI′ **(Conditional copula invariance)** Let $C(\cdot, \cdot; (w, z))$ be the copula for (X, U) $|(W, Z) = (w, z)$ and let $C(\cdot, \cdot; w)$ be the copula for $(X, U)|W = w$. Then $C(\cdot, \cdot; (w, z)) = C(\cdot, \cdot; w)$ for every $(w, z) \in \mathcal{WZ}$.

In Appendix A, I show that all of the identification results adapt naturally to the addition of W under Assumption M$'$. Their omission was only for the sake of notation.

Suppose that for each $k = 1, \ldots, d_x$, $X_k = g_k(W, Z, V_k)$ for some function g_k and random element V_k. V_k may be arbitrarily dependent with U, even after conditioning on W. In fact, this is one motivation for the endogeneity problem studied in this paper since generally if $V_k \not\perp\!\!\!\perp U|W$ then $X_k \not\perp\!\!\!\perp U|W$. The examples in Section 4.2 give several interpretations for the unobservables that can comprise V_k. CCI$'$ is satisfied if Assumption S holds.

Assumption S.

S.1 (Exogenous first stage) $(U, V_k) \perp\!\!\!\perp Z|W$ for every k.

S.2 (Strict monotonicity) For every k, V_k is a scalar and $g_k(w, z, v_k)$ is strictly increasing in v_k for all w, z.

Theorem 2. *Suppose that $X_k = g_k(W, Z, V_k)$ and that Assumptions M.2$'$, M.3$'$ and S hold. Then CCI$'$ is satisfied.*

Theorem 2 extends Theorem 1 of Imbens and Newey (2009).[26] Those authors assume $(U, V_k) \perp\!\!\!\perp (W, Z)$, which is stronger than S.1. As they point out, their assumption requires the instrument to be completely exogenous, rather than just exogenous with respect to the subgroup determined by W. This is an important distinction. In Examples 1, 3 and 4 of Section 4.2, the proposed instruments can reasonably be expected to satisfy S.1, but not the stronger full independence condition. The conclusion of Theorem 1 in Imbens and Newey (2009) is that $U \perp\!\!\!\perp X_k|R_k$ for each k. Proposition 2 shows that this is weaker than the conclusion of Theorem 2.

Many commonly used functional representations of the first stage equation satisfy S.2. Any specification of g that is additively separable in V, such as the classical linear first stage or the nonparametric first stage in Newey et al. (1999) satisfies S.2. However, additive separability is not required, so Theorem 2 also allows for the impact of Z on X to be heterogeneous. In some cases, a structural model may suggest that multiple unobservable factors should enter in g_k and hence that V_k should be a vector, not a scalar. This can be accommodated if v_k is strictly separable from z in g_k.

Definition 2 (Theorem 3.2a, Blackorby et al. (1978)). A vector v_k is strictly separable from z in g_k if and only if there exist scalar-valued functions h_k and $\bar{g}_k$ such that $g_k(w, z, v_k) = \bar{g}_k(w, z, h_k(w, v_k))$ where $\bar{g}_k$ is strictly increasing in h_k for all w, z.[27]

[26] See also Kasy (2011) who considers the relationships between $U \perp\!\!\!\perp X_k|R_k$, $U \perp\!\!\!\perp Z|R_k$ and Theorem 2 when Z is continuous but R_k is a more general control function.

[27] Blackorby et al. (1978) present this definition as a theorem derived from a different definition of strict separability. Equivalence results like this date back to Leontief (1947) and Sono (1947, 1961).

S.2′ (Strict separability) For every k, v_k is strictly separable from z in g_k.

Assumption S.2 can be replaced by the more general S.2′. This follows by defining $\bar{V}_k = h_k(W, V_k)$, for h_k as given in Definition 2. Then S.2 holds with $\bar{V}_k$ in place of V_k and $\bar{g}_k$ in place of g_k. If S.1 holds for V_k, then since h_k does not depend on z, it will also hold for $\bar{V}_k$. Examples 2 and 5 below demonstrate that this slight generalization to strict separability can be quite useful, so I record it for future reference.

Theorem 2′. *Suppose that $X_k = g_k(W, Z, V_k)$ and that Assumptions M.2′, M.3′, S.1 and S.2′ hold. Then CCI′ is satisfied.*

An alternative set of sufficient conditions can be stated using counterfactual notation. For every $z \in \mathcal{Z}$, let $X(z)$ be the value that X would obtain if Z were exogenously (or counterfactually) set to z. Then $X(z)$ is observable as X if and only if $Z = z$, i.e. $X = \sum_{z \in \mathcal{Z}} \mathbb{1}[Z = z] X(z)$, where $\mathbb{1}$ is the indicator function. When $Z = z$, the counterfactual outcome $X(z')$ is unobservable for $z' \neq z$. This construction is especially useful when X is thought of as being chosen by an agent who knows their realization of Z. It allows the analyst to mentally separate the choice-generated dependence between observed outcomes X and Z from the intrinsic, choice-invariant dependence between counterfactual outcomes $X(z)$ and $X(z')$. Note that in general, $X(z)$ is still random since other observable and unobservable factors besides Z play a role in the determination of X.

Assumption C.

C.1 (Weak unconfoundedness of the instrument) $(X_k(z), U) \perp\!\!\!\perp Z | W$ for all $z \in \mathcal{Z}$ and every k.

C.2 (First stage rank invariance) $F_{X_k(z)|W}(X_k(z) \mid w) \overset{\text{a.s.}}{=} F_{X_k(z')|W}(X_k(z') \mid w)$ for all $(w, z), (w, z') \in \mathcal{W}\mathcal{Z}$ and every k.

Theorem 3. *Suppose Assumptions M.2′, M.3′ and C are satisfied. Then Assumption CCI′ is satisfied.*

Assumption C.1 is familiar from the literature on local average treatment effects initiated by Imbens and Angrist (1994). It says that, conditional on covariates, the instrument acts as an exogenous treatment on X. It is nearly equivalent in content (although technically weaker) than its counterpart S.1. Assumption C.2 is an old assumption that is originally due to Doksum (1974). Its interpretation is that agents possess some underlying latent characteristic that determines their relative outcome of X under each counterfactual outcome of Z.

Rank invariance can be a strong assumption, however it is often held implicitly in modeling frameworks that do not use counterfactual notation. For example, it is implied by S.2 since there the counterfactual outcome of X_k is $X_k(z) = g_k(W, z, V)$ and thus by monotonicity, $F_{X_k(z)|W}(X_k(z) \mid w) = F_{V|W}(V \mid w) = F_{X_k(z')|W}(X_k(z') \mid w)$. As noted, many commonly used econometric models satisfy S.2. It is also implied with regards to counterfactual outcomes of Y by M.1′. As discussed by Heckman et al. (1997), even an ideal randomized experiment does not shed light on the joint distribution of $\{X(z)\}_{z\in\mathcal{Z}}$ since by definition only one of $X(z)$ is ever observed. For the same reason, any assumption about the joint distribution of $\{X(z)\}_{z\in\mathcal{Z}}$ is inherently untestable. Chernozhukov and Hansen (2005) consider a slightly more general assumption that allows the ranks $F_{X_k(z)|W}(X_k(z) \mid w)$ and $F_{X_k(z')|W}(X_k(z') \mid w)$ to deviate idiosyncratically from some common rank. This has the same implication for the observed data as C.2 and does not allow for systematic differences in rank over counterfactual states. Meaningful alternatives to rank invariance can come from additional assumptions on behavior, such as those discussed by Heckman et al. (1997). A prominent example of such a behavioral assumption is the instrument monotonicity condition of Imbens and Angrist (1994), which requires that either $X(z) \geq X(z')$ a.s. or $X(z) \leq X(z')$ a.s. for all $z, z' \in \mathcal{Z}$.[28] Note that monotonicity and C.2 are not nested.

4.2 Examples

Example 1 (**The causal effect of class size**)**.** Let Y be a measure of schooling outcomes (e.g., standardized test scores), X be the average class size of a school and W be a set of observable controls such as school characteristics and socioeconomic variables. The unobservable U aggregates the litany of other factors involved in determining outcomes, including parental involvement and unobserved family background characteristics. The assumption that $U \perp\!\!\!\perp X$ is untenable, even conditional on W, because families that value education more highly are more likely to select into schools on the basis of the prevailing class size.

Feinstein and Symons (1999) use geographic indicator variables for instruments. Variation in these indicators corresponds to different local authorities (a unit of local government in England) which have different policies on class size. The authors cite work on the determinants of migration to argue that geographic location at the local authority level is exogenous to schooling outcomes after conditioning on measures of social class, parents' education and parental interest. That is, C.1 holds when W is a

[28] For careful discussions of this condition, see the subsequent papers by Angrist and Imbens (1995), Angrist et al. (1996) and Vytlacil (2002).

set of controls containing these variables.

To assess the plausibility of C.2, consider two subpopulations, $P_0 : (W, Z) = (w, 0)$ and $P_1 : (W, Z) = (w, 1)$, corresponding to two different local authorities. Rank invariance requires that a school in P_0 that has a small class size relative to other schools in P_0 would also have a small class size relative to schools in P_1 if it were, counterfactually, a member of P_1. In other words, whatever factor it is that makes a school have a small class size among other schools in P_0 is intrinsic to the school and does not interact with the geographic region. Note that the absolute class size of the school can be different under counterfactual assignment to P_0 and P_1. In fact, the absolute class size must be affected by the instrument if it is to satisfy Assumption R. Rank invariance is reasonable in this application given the included control set. The remaining variation in class size within P_0 may be due to, for example, an unusually pushy set of parents. Rank invariance is the assumption that the same parents would still be pushy if the school were counterfactually located in a different local authority. ■

Example 2 (**The causal effect of class size with a first stage equation**)**.** Consider the same inference problem as in Example 1. Hoxby (2000) uses an instrument that captures the exogenous fluctuations in the number of enrolled students caused by changes in the timing of births around the calendar year. Suppose that $\bar{s}(W, V)$ represents the number of students that would be enrolled if the timing of births were non-varying, where V is some random element that may be arbitrarily dependent with U. Letting $Z \geq 0$ represent proportional exogenous fluctuations in enrollment, the actual number of enrolled students is $s(W, Z, V) = Z\bar{s}(W, V)$. If $\bar{c}(W, V)$ denotes the number of classes then

$$X = \frac{s(W, Z, V)}{\bar{c}(W, V)} = Z\frac{\bar{s}(W, V)}{\bar{c}(W, V)} = Zh(W, V). \tag{12}$$

Thus $X = g(W, Z, V) = Zh(W, V)$ is strictly separable in V and satisfies S.2′. Under the assumption that $(U, V) \perp\!\!\!\perp Z|W$, CCI′ holds by Theorem 2′.

An objection to this formulation is that a school may have advance knowledge of Z and adjust the number of classes accordingly. In this case, the number of classes should also depend on Z. This can be addressed by specifying the number of classes as $c(W, Z, V) = d(W, Z)\bar{c}(W, V)$, where $d \geq 0, d(W, 1) = 1$ is an adjustment multiplier. Then (12) becomes

$$X = \frac{Z}{d(W, Z)}\frac{\bar{s}(W, V)}{\bar{c}(W, V)} = h^z(W, Z)h(W, V),$$

which still satisfies S.2′ with $X = g(W, Z, V) = h^z(W, Z)h(W, V)$. CCI′ would still hold in this case.

It could also be argued that the way in which a school adjusts the number of classes in response to an exogenous enrollment shock should depend on V, the unobservable which may be dependent with U. For example, one might think that even after conditioning on covariates, schools with high levels of parental involvement are more responsive to exogenous shocks than other schools. Assuming that parental involvement is unobservable and cannot be properly proxied for, its effect will be captured in V. In this case, d should be a function of W, Z and also V. Assumption S.2′ no longer holds and consequently CCI′ may not either.

Intuitively, the reason that CCI′ might not hold in this latter case is that the exogenous fluctuations, Z, could themselves carry additional information on the nature of the selection problem that could not be controlled for. For example, suppose that parents select into schools not only on the basis of class size, X, but also with a preference for how the school responds to Z, a characteristic controlled by the function d, which I am now allowing to depend on V. By itself, X can carry some information about V, but when combined with Z, it could potentially carry more. The purpose of S.2′ is to ensure that it does not, after one controls for the marginal effect that Z has on X, which is observable. Without such an assumption (or more structure), it is not clear how one can make any causal shift in Z while conditioning on X, since Z would provide additional information on V and V is generally dependent with U. In other words, the unobservable characteristics of a school with any given (X, Z) combination could be entirely unlike the characteristics of any other school in the population. Adding Z only serves to further complicate the original selection problem, even though $U \perp\!\!\!\perp Z$.

Separable models can ignore this issue by making the strong assumption that there is one causal effect of class size that is applicable to all observationally equivalent schools. The possibility that Z could provide additional information on U, via V, after conditioning on X, is completely irrelevant in a separable model because the causal effect of class size is assumed to not depend on U. Given the selection problem and the likelihood of heterogeneity in the marginal effect of class size, CCI′ may still be a relatively unobjectionable assumption, even given the preceding discussion. ■

Example 3 (**Estimating returns to schooling with a natural experiment**)**.** Duflo (2001) estimates the returns to schooling using an instrument motivated by a natural experiment in Indonesia. The experiment was caused by a 1970's era government campaign to construct primary schools. The author shows that attained education was higher among children who were more strongly exposed to this program due to being

younger and living in districts where more schools were built. In this example, Y is hourly wage, X is schooling obtained and W contains characteristics of the region of birth. The instrument, Z, is an interaction term between year of birth and the number of schools planned for construction in the agent's region of birth. Duflo argues convincingly that the instrument can be excluded from (1) and that C.1 holds. Assumption C.2 requires that agents who would have obtained a large amount of schooling, relative to their peers, had they been of an age and in a region where many schools were built, would also have obtained a relatively large amount of schooling if they were of an age/region in which few schools were built. In other words, C.2 assumes that the agents possess an underlying proclivity for education that does not interact with their year of birth or the intensity of the school building program in their birth region.

■

Example 4 (A structural model of the returns to schooling[29]). Suppose that X is a measure of investment in schooling, Y is lifetime earnings and W is a set of socioeconomic and family background controls. The analyst is interested in the causal effect of X in the educational production function $Y = m(X, W, U)$. The classic endogeneity problem in this situation is that $X \not\perp\!\!\!\perp U|W$ because U captures, among other things, latent traits such as ability, which are expected to be dependent with both education decisions and earnings.

Suppose that agents choose X by maximizing expected lifetime earnings net of costs,

$$X = g(W, Z, V) = \arg\max_x \mathbb{E}\left[m(x, W, U)|V, W\right] - c(x, Z, W),$$

where c is the educational cost function and V is a scalar signal of U that is observed by the agent. The instrument Z is an exogenous cost shifter that is excluded from the production function. For example, Card (1995) uses an indicator for living in a county with a four-year college as a cost-shifter. This instrument is unlikely to satisfy the full independence condition $(U, V) \perp\!\!\!\perp (W, Z)$, or even $(U, V) \perp\!\!\!\perp Z$, since family background characteristics are likely to be dependent with both ability and proximity to a college. However, after conditioning on these observed characteristics, Z is plausibly exogenous and S.1 is reasonable.

As discussed in Imbens and Newey (2009), g is strictly increasing in v under the following assumptions.

[29] This example is due to Imbens and Newey (2009). I have modified it slightly to highlight the importance of covariates.

1. m is strictly increasing in X.
2. m is twice continuously differentiable.
3. There are diminishing returns to schooling, $\nabla_x^2 m < 0$, and the returns to schooling increase in ability, $\nabla_{x,u}^2 m > 0$.
4. Costs increase in education at an increasing rate, $\nabla_x c, \nabla_x^2 c > 0$.
5. U and V are affiliated random variables, conditional on W.

CCI′ is then satisfied by Theorem 2. ■

Example 5 (**Nonseparable supply and demand**)**.** Identification in nonseparable simultaneous equations models has been studied recently by Benkard and Berry (2006), Matzkin (2008) and Blundell and Matzkin (2010). A case of particular historical interest is the demand system

$$\begin{aligned} Q &= d(P, U^d), \\ P &= s(Q, Z, U^s), \end{aligned} \tag{13}$$

where Q is quantity, P is price and U^s, U^d are scalar, latent supply and demand shocks. The instrument Z is an observed supply shifter that can be excluded from the demand equation. A classic example of Z is a weather shock in a market for an agricultural product. The analyst is interested in identifying the demand function, d, when M.1 and M.2 are satisfied. The following assumptions are sufficient for M.3 and CCI.

Assumption D.

D.1 (Smoothness) d and s are twice continuously differentiable.

D.2 (Convex support) The support of (Q, P) is convex.

D.3 (Monotonicity of supply) s is strictly increasing in u^s.

D.4 (Demand slopes downward, supply slopes upward) d is strictly decreasing in p and s is strictly increasing in q.

D.5 (Exogenous instrument) $Z \perp\!\!\!\perp (U^d, U^s)$.

D.6 (Monotonic instrument) s is strictly increasing in z.

D.7 (Strictly separable supply) z is strictly separable from (q, u^s) in s.

Proposition 6. *Suppose Q, P, U^d and d satisfy M.1 and M.2 and that Assumption D is satisfied. Then Assumptions M.3 and CCI hold.*

D.1 and D.2 are regularity conditions needed to ensure the existence of a smooth global reduced form function. D.6 is similar to the monotonicity condition of Imbens and Angrist (1994). Whether it is reasonable or not depends on the context. An example where it is reasonable is given by Angrist et al. (2000), who use weather variables as exogenous instruments in a fish market. It also seems appropriate in the case when Z is rainfall and the market is for an agricultural good. D.7 is a general restriction on the functional form of the supply curve. For example, it is satisfied if $s(q, z, u^s) = s^Z(z)\bar{s}(q, u^s)$ for any functions $s^Z > 0$ and $\bar{s}$ that are strictly increasing in each argument.

Blundell and Matzkin (2010) study simultaneous systems that satisfy a subset of Assumption D. In particular, Blundell and Matzkin (2010) do not use D.6 or D.7. They show in their Theorem 1 that such a system is observationally equivalent to a triangular system satisfying Assumptions M and S if and only if the inverse of s with respect to u^s adheres to a specific functional structure which they call control function separability. An implication of Proposition 6 is that the low-level conditions in D.6 and D.7 are sufficient for control function separability—see the proof for more detail. It follows that Assumption D must be stronger than needed to secure M.3 and CCI and hence identification of d.

■

5 Estimation

Estimation for the case when $d_x = 1$ can be based on the criterion function $T : \mathcal{M} \to \mathbb{R} : T(m) = \mathbb{E}\,\Gamma(m, X, Z)$, where $\mathcal{M}$ is the collection of all functions that satisfy Assumption M,

$$\Gamma(m, x, z) = \left\| Q_{Y|X}\left(F_{Y|XZ}\left(m\left[q(x|z), \cdot\right] \mid q(x|z), z\right) \mid x\right) - m(x, \cdot)\right\|_U^2, \qquad (14)$$

$q(x|z) = Q_{X|Z}(F_X(x) \mid z)$, and $\|\cdot\|_U$ is the L_2-norm with respect to Lebesgue measure on the unit interval. This is a redefinition of T and Γ from Section 3, where the criterion function was taken to be a functional of the copula. The arguments in that section show that the two approaches are equivalent and that $T(m) = 0$ if and only if $m = m_0$, where m_0 is the "true" element of $\mathcal{M}$ that is assumed to generate the observed data through (1). These notational changes will cause no confusion in the following. Directly estimating m_0 is useful because it may be natural to specify a parametric or semiparametric form for the collection $\mathcal{M}$, whereas this is less likely to be the case for $\mathcal{C}$.

I make some simplifying assumptions in this section. First, for readability, I omit covariates and impose Normalization N.1. The estimators and assumptions adjust in the expected ways when there are covariates. I make the restrictive assumption, which is already reflected in (14), that there is only one endogenous variable, i.e. $d_x = 1$. If $d_x > 1$ and the components of X are not assumed to be mutually independent, then the scaling factor σ, defined in Proposition 3 and 4, must also be estimated. This does not cause any significant complications in the asymptotic theory.

A natural estimator of m_0 is the minimizer over $\mathcal{M}$ of the sample analog of T, i.e. the sample average of Γ. However, this estimator needs to be adjusted in two ways. First, if $\mathcal{M}$ is infinite-dimensional, optimizing over $\mathcal{M}$ would be practically infeasible and, even if it were not, would typically produce an estimator with poor asymptotic properties. An attractive solution is the method of sieves, which replaces optimization over $\mathcal{M}$ with optimization over a finite-dimensional subset, $\mathcal{M}_n$, that grows with the sample size. This makes implementation feasible and also allows for finer control of the asymptotic behavior of the estimator.[30]

A second adjustment is needed because Γ depends on features of the distribution of (Y, X, Z) and so must be estimated. As an estimator of Γ I take

$$\hat{\Gamma}(m, x, z) = \left\| \hat{Q}_{Y|X} \left(\hat{F}_{Y|XZ} \left(m \left[\hat{q}(x|z), \cdot \right] \mid \hat{q}(x|z), z \right) \,\middle|\, x \right) - m(x, \cdot) \right\|_U^2,$$

where $\hat{q}(x|z) = \hat{Q}_{\bar{X}|Z}(\hat{F}_{\bar{X}}(x) \mid z)$. The random variable $\bar{X} = X|[X \in \overline{\mathcal{X}}]$, which I will discuss more in what follows, is a truncation of X to a compact interval, $\overline{\mathcal{X}}$, that is strictly contained in $\mathcal{X}$. The estimators $\hat{Q}_{Y|X}, \hat{F}_{Y|XZ}, \hat{Q}_{\bar{X}|Z}$ and $\hat{F}_{\bar{X}}$ are kernel-based estimators defined as follows:

$$\begin{aligned}
\hat{F}_{Y|XZ}(y \mid x, z) &= \frac{\sum_{i=1}^n \mathbb{1}[Y_i \le y] K_i^{xz}}{\sum_{i=1}^n K_i^{xz}}, && \text{and } K_i^{xz} = K\left(\frac{(x,z) - (X_i, Z_i)}{h_{xz}} \right); \\
\hat{Q}_{Y|X}(t \mid x) &= \inf\{y : \hat{F}_{Y|X}(y \mid x) \ge t\}, \text{ where} \\
\hat{F}_{Y|X}(y \mid x) &= \frac{\sum_{i=1}^n \mathbb{1}[Y_i \le y] K_i^{x}}{\sum_{i=1}^n K_i^{x}}, && \text{and } K_i^{x} = K\left(\frac{x - X_i}{h_{x}} \right); \\
\hat{Q}_{\bar{X}|Z}(t \mid z) &= \inf\{x : \hat{F}_{\bar{X}|Z}(x \mid z) \ge t\}, \text{ where} \\
\hat{F}_{\bar{X}|Z}(x \mid z) &= \frac{\sum_{i \in \mathcal{I}_{\overline{\mathcal{X}}}} \mathbb{1}[X_i \le x] K_i^{z}}{\sum_{i \in \mathcal{I}_{\overline{\mathcal{X}}}} K_i^{z}}, && \text{and } K_i^{z} = K\left(\frac{z - Z_i}{h_{z}} \right); \\
\text{and } \hat{F}_{\bar{X}}(x) &= \frac{1}{|\mathcal{I}_{\overline{\mathcal{X}}}|} \sum_{i \in \mathcal{I}_{\overline{\mathcal{X}}}} \mathbb{1}[X_i \le x].
\end{aligned}$$

[30] See Chen (2007) for a complete survey of the method of sieves.

In these definitions K is (as appropriate) a $1, d_z$ or (d_z+1)-dimensional kernel function with standard properties to be specified in Assumption E.5, and $\mathcal{I}_{\overline{\mathcal{X}}} = \{i : X_i \in \overline{\mathcal{X}}\}$ indexes the observations that fall in $\overline{\mathcal{X}}$. The feasible sieve extremum estimator, $\hat{m}$, is then given by

$$\hat{m} = \underset{m \in \mathcal{M}_n}{\arg\min} \frac{1}{|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}|} \sum_{i \in \mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \hat{\Gamma}(m, X_i, Z_i), \tag{15}$$

where $\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}} = \{i : (X_i, Z_i) \in \overline{\mathcal{X}\mathcal{Z}}\}$ indexes the observations that fall in $\overline{\mathcal{X}\mathcal{Z}}$, a set that is defined in E.2 and discussed more below. Let $\|m\|_\infty = \sup_{(x,u)\in\overline{\mathcal{X}}\times[0,1]} |m(x,u)|$ denote the sup-norm on $\overline{\mathcal{X}} \times [0,1]$. The following conditions are sufficient for $\hat{m}$ to be strongly consistent for m_0.

Assumption E.

E.1 (Random sample) $(Y_i, X_i, Z_i), i = 1, \ldots, n$ are independent and identically distributed.

E.2 (Compact support) If $\mathcal{Z}$ is not finite then $\overline{\mathcal{X}}$ and $\overline{\mathcal{Z}}$ are compact sets such that $\overline{\mathcal{X}} \times \overline{\mathcal{Z}}$ is strictly contained in $\mathcal{X}\mathcal{Z}$. If $\mathcal{Z}$ is finite then $\overline{\mathcal{X}}$ is a compact set that is strictly contained in $\mathcal{X}$ and $\overline{\mathcal{Z}} = \mathcal{Z}$. Write $\bar{X} = X|[X \in \overline{\mathcal{X}}]$, $\bar{Z} = Z|[Z \in \overline{\mathcal{Z}}]$ and let $\overline{\mathcal{X}\mathcal{Z}}$ denote the support of $(\bar{X}, \bar{Z})$.

E.3 (Smooth parameter space) $\mathcal{M}$ is uniformly Lipschitz, i.e. there exists a constant $B > 0$ such that $|m(x,u)| \leq B$ and $|m(x,u) - m(\tilde{x},\tilde{u})| \leq B \|(x,u) - (\tilde{x},\tilde{u})\|$ for every $m \in \mathcal{M}$, every $(x,u), (\tilde{x},\tilde{u}) \in \overline{\mathcal{X}} \times [0,1]$.

E.4 (Smooth, well-behaved distribution functions) $F_{Y|XZ}$ is continuously differentiable in (x,z) on $\overline{\mathcal{X}}\times\overline{\mathcal{Z}}$ (or just in x if $\mathcal{Z}$ is finite); f_{XZ} and f_Z are continuous (just in x if $\mathcal{Z}$ is finite) and bounded away from 0 on $\overline{\mathcal{X}}\times\overline{\mathcal{Z}}$ and $\overline{\mathcal{Z}}$. Also, $f_{Y|XZ}$ is bounded uniformly in (y,x,z) on the support of $(Y, \bar{X}, \bar{Z})$, denoted as $\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}$, and $f_{Y|X}$ is bounded away from 0 uniformly in (y,x) on $\mathcal{Y}\overline{\mathcal{X}}$, the support of $(Y, \bar{X})$.

E.5 (Kernels and bandwidths) K is (as appropriate) a $1, d_z$ or (d_z+1)-dimensional compactly supported second order kernel function that is bounded and symmetric. As $n \to \infty$, h_{xz}, h_z and $h_z \to 0$, and $n^{1+d_z}(\log n)^{-1}h_{xz}$, $n^{d_z}(\log n)^{-1}h_z$ and $n(\log n)^{-1}h_x \to \infty$.

E.6 (Sieve spaces) $\{\mathcal{M}_n\}_{n=1}^\infty$ are finite-dimensional compact subsets of $\mathcal{M}$ with $k_n = \dim \mathcal{M}_n \to \infty$. Choose $\{\mathcal{M}_n\}_{n=1}^\infty$ such that for any $m \in \mathcal{M}$, there exists a sequence $\{\Pi_n m \in \mathcal{M}_n\}_{n=1}^\infty$ satisfying $\|\Pi_n m - m\|_\infty = o(1)$.

Theorem 4. *Suppose Assumption M holds for* $(Y, \bar{X}, \bar{Z})$ *and that Assumption R holds almost everywhere for* $(\bar{X}, \bar{Z})$*, i.e.* $\bar{X} \not\perp\!\!\!\perp \bar{Z}$ *at almost every* $x \in \bar{X}$*. Then under E,* $\|\hat{m} - m_0\|_\infty = o_{\mathbb{P}}(1)$*.*

The proof of this theorem divides the estimation problem into two sub-problems. The first is to uniformly estimate Γ by $\hat{\Gamma}$. The second problem is the infeasible sieve estimation problem with Γ instead of $\hat{\Gamma}$. Assumption E reflects the combined assumptions from these two different estimation problems.

Assumptions E.1-E.5 are standard for uniformly estimating a conditional distribution function with a Nadaraya-Watson kernel estimator. The proof of Theorem 4 depends on the uniformity of this estimator over all of its arguments. This was originally shown by Härdle et al. (1988) and has been strengthened considerably by Einmahl and Mason (2000, 2005). The assumptions on the kernel function and the bandwidth in E.5 are quite general and, as the previous authors show, can be made even weaker than I have presented them. While a formal analysis of data-driven bandwidth selection is far beyond the scope of this paper, I implement a cross-validation approach in the simulations in Section 6. I note here that the results of Einmahl and Mason (2005) are also uniform over bandwidth and so immediately imply that Theorem 4 still holds when cross-validating.

Generally, it is difficult to uniformly estimate a function with kernel regression over a non-compact set. In many contexts, it is reasonable to restrict inference to some compact subset $\overline{\mathcal{XZ}} \subsetneq \mathcal{XZ}$ and in many estimation problems this can be done without further qualification. However, in this model, identification depends on $\Gamma(\cdot, x, z) = 0$ for a.e. $(x, z) \in \mathcal{XZ}$, so just arbitrarily choosing some $\overline{\mathcal{XZ}}$ could lead to a loss of identification. The researcher must therefore ensure two things when choosing the estimation region $\overline{\mathcal{XZ}}$. First, that Assumption R holds over $\bar{X}$. Second, that CCI holds for $(\bar{X}, \bar{Z})$. In general, this is not ensured by CCI holding for (X, Z). In most cases, however, this should not present a serious problem, since justifying CCI forces the analyst to consider the relevant support. If this justification is through specifying a first stage equation, the validity of a proposed $\overline{\mathcal{XZ}}$ will be immediately clear. In addition, Proposition 5 shows that for a particular set, $\mathcal{X}^D$, defined as the collection of all points of local dependence, both Assumptions R and CCI are always satisfied. The uniformity issue is resolved by using (X, Z) to estimate the conditional distribution functions, but using only $(\bar{X}, \bar{Z})$ to compute the sample average (15).

The infeasible sieve estimation problem is an application of the theory of White and Wooldridge (1991) and Chen (2007) for approximate sieve M-estimators. This part of the problem depends on E.1-E.4 and E.6. Assumption E.3 ensures that functions in

$\mathcal{M}$ are sufficiently smooth and, together with E.2, also makes $(\mathcal{M}, \|\cdot\|_\infty)$ compact.

The finite-dimensional sets $\{\mathcal{M}_n\}_{n=1}^\infty$ in E.6 are called sieve spaces and are used to approximate $\mathcal{M}$. This approximation introduces a deterministic bias into the estimator that is eliminated asymptotically by choosing $\{\mathcal{M}_n\}_{n=1}^\infty$ to be dense in $\mathcal{M}$ and letting $k_n = \dim \mathcal{M}_n \to \infty$. A common approach is to take $\mathcal{M}_n$ to be a linear basis, a simple example of which is the polynomial (power series) basis:

$$\mathcal{M}_n = \left\{ m(x,u) = \sum_{l=0}^{k_n^u - 1} \sum_{j=0}^{k_n^x - 1} \theta_{j,l} x^j u^l : \theta \in \Theta \subseteq \mathbb{R}^{k_n^x + k_n^u} \right\},$$

where Θ is a compact set and $k_n = k_n^x + k_n^u$. Other examples of linear bases are polynomial splines and Hermite polynomials. There are also nonlinear bases such as neural networks and E.6 allows for these as well. Many of these bases are particularly well-suited for approximating functions with certain smoothness or shape properties, and these considerations inform the choice of basis in practice.[31] Finally, I note that the sieve estimation framework nests parametric specifications of $\mathcal{M}$ (i.e. $\dim \mathcal{M} < \infty$) by taking $\mathcal{M}_n = \mathcal{M}$ for all n.

6 Monte Carlo

A surprising aspect of the identification result of this paper is that it allows for discrete instruments. This is unusual in the literature on nonseparable models with continuous endogenous explanatory variables. As a result, one might suspect that discrete instruments represent a sort of limiting case that may perform quite poorly in practice. In this section I report the results of Monte Carlo simulations which show that this is not the case.

The simulations use one of the following parametric specifications of the outcome equation,

$$Y = \theta(X - 10) + XU/10, \tag{MC1}$$

$$Y = \theta(X - 10)^2 + XU/10, \tag{MC2}$$

$$Y = \theta(X - 10)^2 + \theta^2(X - 10) + XU/10, \tag{MC3}$$

where θ is the unknown parameter and I assume that the functional form is known. All of these models are nonseparable and, although parametric, it is not clear how they

[31] See Section 2.3 of Chen (2007) for a survey of different sieve bases and their approximation properties.

could be estimated without the results of this paper, given that the distribution of $U|X$ is completely unknown. The significance of the number 10 in (MC1)-(MC3) is to impose Normalization N.2 in Appendix B, i.e. $m(10, U; \theta) = U$ for all θ. Specification (MC1) is linear in both X and θ, while (MC2) is nonlinear in X, but linear in θ and (MC3) is nonlinear in both X and θ.

I generated samples of size $N = 200, 400$ and 800 for $X \sim N(10 + \mu_Z Z, 1)$, $U \sim N(0, 1)$ and (X, U) linked together by the Clayton copula $C(r, u; \rho) = (r^{-\rho} + u^{-\rho} - 1)^{-1/\rho}$, with $\rho = 1$. The data was generated with $\theta_0 = 1$ for (MC1), $\theta_0 = .3$ for (MC2) and $\theta_0 = .5$ for (MC3). I used values of $1, .5$ and $.25$ for μ_Z, which controls the relevance of the instrument for X. For each configuration, I considered three different marginal distributions for Z: Bernoulli (binary), uniform and normal, all chosen to have mean $1/2$ and variance $1/4$. Estimation followed (15) with a bi-weight kernel. I used least-squares cross-validation to compute the bandwidths for the estimators of the conditional distribution functions.

The results are presented in Tables 1-3, each of which corresponds to an outcome equation (MC1)-(MC3). Each table reports the approximate finite-sample bias, standard deviation and mean squared error (MSE) for each of the 27 combinations of μ_Z, the marginal distribution of Z and the sample size, N. The approximations are based on 500 replications. The results for some selected simulations were insensitive to higher numbers of replications. The estimator seems to perform fairly well, except in particularly unfavorable cases with a poor instrument and a small sample size. As expected, the MSE is decreasing in instrument relevance, μ_Z, and increasing in N. The bias and standard deviation do not always have this monotone pattern, but this is not unusual since the estimator involves kernel smoothing. I hesitate to draw any sweeping conclusions on the performance of the estimator since there are many dimensions on which these simulations can be adjusted. These include the marginal distribution of X, the way in which Z affects X, and the severity and form of the dependence between X and U.

However, one important conclusion that can be confidently drawn from the simulation results is that the binary instrument performs as well, if not better, than either of the continuous instruments. This was not clear from the outset. Since the proof of Theorem 1 relies on continuity and sequencing arguments, one might reasonably have suspected that the identification argument works poorly in practice. It is difficult to form an exact comparison between the binary and continuous instruments, however. This is because the rate of convergence of the estimator is probably limited by the rate of convergence of $\hat{F}_{Y|XZ}$, which in turn is determined by the dimension of (X, Z). When Z is discrete, this dimension is smaller since the data is being "binned" rather

than smoothed. Discrete instruments thus have an inherent advantage in estimation, which could actually offset a disadvantage in identifying content. The conservative conclusion is that discrete and continuous instruments perform comparably.

Finally, for illustrative purposes, I include the correlation coefficient between X and Z in Tables 1-3, because this is a standard measure of the relevance of an instrument in applied work. It is interesting to observe that it is a poor predictor of how well an instrument performs in this model. For example, in Table 1 with $\mu_Z = .5$, all marginal distributions of Z generate similar correlation coefficients, but the normally distributed Z performs markedly worse. Correlation is a measure of linear dependence, so the simulation results show that the model is able to exploit the nonlinear dependence between X and Z in a way that is not reflected by the linear correlation. This is not surprising in light of the theory presented in this paper, but it demonstrates a departure from the standard intuition for instrumental variable methods.

7 Conclusion

In this paper I have discussed an assumption called conditional copula invariance (CCI). I have shown that CCI can be used to identify a general nonparametric, nonseparable instrumental variable model. An unusual aspect of the identification result is that it holds for nearly any reasonable instrument. Since discrete instruments are widely used in applications and good (i.e., plausibly exogenous) instruments of all types are rare, this can be important in practice. Another practical importance of CCI is that it can be clearly interpreted in economic terms, which makes inference based on it more transparent.

I analyzed a sieve extremum estimator and proved its strong consistency. In a semiparametric Monte Carlo simulation, I showed that it performs fairly well in relatively small samples with either discrete or continuous instruments. The estimation procedure in this paper can be further developed in several ways using the general theory of sieve M-estimation. These include results on the rate of convergence, the asymptotic normality of finite-dimensional features of m and various dimension reduction strategies. I am pursuing these topics in a companion paper.

A Covariates

In this appendix I indicate how the results adjust to accommodate for covariates. Note that Assumption M$'$ was given in Section 4. Denote the marginal support of W by

$\mathcal{W} \subseteq \mathbb{R}^{d_w}$ and modify (1) to be $Y = m(X, W, U)$. Also let $\mathcal{XW}, \mathcal{WZ}$, and $\mathcal{XWZ}$ be the joint supports of $(X, W), (W, Z)$ and (X, W, Z), respectively. M.1′-M.3′ and CCI′ are natural extensions of their counterparts in Assumption M for the general case when W are possibly endogenous, i.e. $U \not\perp\!\!\!\perp W$. Propositions 1 and 2 change little after redefining R_k as $R_k = F_{X_k|WZ}(X_k \mid W, Z)$.

Proposition 1′. *Given M.2′-M.3′, CCI′ holds if and only if for every k, $U \perp\!\!\!\perp Z | W, R_k$.*

Proposition 2′. *Given M.2′-M.3′, CCI′ implies that for every k, $U \perp\!\!\!\perp X_k | W, R_k$.*

Likewise, Propositions 3 and 4 change slightly and in the same way after redefining $\bar{m}(x, w, s) = m(x, w, Q_{U|W}(s \mid w))$.

Proposition 3′ and 4′. *Under M.1′ and M.2′,*

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), s; w\right) = F_{Y|XW}\left(\bar{m}(x, w, s) \mid x, w\right)/\sigma(x|w) \text{ and}$$
$$\bar{m}(x, w, s) = Q_{Y|XW}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), s; w\right)\sigma(x|w) \,\middle|\, x\right)$$

for $(x, w) \in \mathcal{XW}, s \in (0, 1)$ where $\sigma(x|w) = \left(\prod_{k=1}^{d_x} f_{X_k|W}(x_k \mid w)\right)/f_{X|W}(x \mid w)$. If in addition M.3′ and CCI′ hold then

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}(x \mid w, z), s; w\right) = F_{Y|XWZ}\left(\bar{m}(x, w, s) \mid x, w, z\right)/\sigma(x|w, z) \text{ and}$$
$$\bar{m}(x, w, s) = Q_{Y|XWZ}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}(x \mid w, z), s; w\right)\sigma(x|w, z) \,\middle|\, x, w, z\right),$$

with $\sigma(x|w, z) = \left(\prod_{k=1}^{d_x} f_{X_k|WZ}(x_k \mid w, z)\right)/f_{X|WZ}(x \mid w, z)$.

These corollaries lead to an observable implication analogous to (4),

$$\begin{aligned} & Q_{Y|XW}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), s; w\right)\sigma(x|w) \,\middle|\, x, w\right) \\ = \; & Q_{Y|XWZ}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}(x \mid w, z), s; w\right)\sigma(x|w, z) \,\middle|\, x, w, z\right), \end{aligned} \tag{16}$$

from which the main identification result follows with a slight rephrasing of the rank condition.

Theorem 1′. *Redefine Γ and T in the natural way to account for the inclusion of W.*[32] *Assume that $X \not\perp\!\!\!\perp Z | W = w$ at x, i.e. $\mathbb{P}\{z \in Z : F_{X|WZ}(x \mid w, z) \neq F_{X|W}(x \mid w)\} > 0$. Then under Assumption M′, $C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), s; w\right)$ is identified for every $s \in [0, 1]$. Thus $\bar{m}(x, w, s)$ is identified.*

[32] How this is done should be clear from comparing (4) with (16).

The interpretation of the identified object now depends on whether or not M.3′ can be strengthened by the additional assumption that $U \perp\!\!\!\perp W$. If it cannot, then the identified quantity is $\bar{m}(x, w, s) = m(x, w, Q_{U|W}(s \mid w))$, which is the s^{th} quantile of outcomes for the subpopulation $W = w$ when X is exogenously set to x. The causal effect of X is identified, but the causal effect of belonging to the subpopulation defined by $W = w$ is not. This should be expected since the model with M.3′ places no restrictions on the dependence between U and W. This is still sufficient for useful inference if the purpose of the analysis is to study the causal effect of X. If M.3′ is strengthened to $U \perp\!\!\!\perp W$ then the causal effects of both X and W are identified.

B Alternative Normalizations

As mentioned, N.1 is one of several normalizations that can be used to constructively identify m from $\bar{m}$. In this appendix, I discuss some others. However, I note that even without a normalization, $\bar{m}(x, s) = m(x, Q_U(s))$ is always interpretable as the s^{th}-QTR to $X = x$. In addition, as long as U is continuously distributed, one can draw an independent $S \sim \text{Unif}[0, 1]$ and generate $Q_U(S) \sim U$ and $\bar{m}(x, S) = m(x, Q_U(S)) \sim m(x, U)$. For example, identification of the average treatment effect does not require a normalization since $\mathbb{E}\,\bar{m}(x, S) = \mathbb{E}\,m(x, U)$.

The following alternative normalizations were developed by Matzkin (2003).

Proposition B.1. *If $\bar{m}(x, w, s)$ is identified then $m(x, w, u)$ is constructively identified under any of N.2-N.4.*[33]

N.2 (Scale) There is a known $\bar{x} \in \mathcal{X}$ such that $m(\bar{x}, w, u) = u$ for all u. In particular, this yields

$$m(x, w, u) = \bar{m}(x, w, \bar{m}^{-1}(\bar{x}, w, u)).$$

N.3 (Homogeneity of degree one) For a known $\bar{x}, \bar{u} > 0, \alpha > 0$ and for all $\lambda > 0$, $m(\bar{x}, w, \lambda\bar{u}) = \lambda\alpha$ for all w. In this case,

$$m(x, w, u) = \bar{m}\left(x, w, \bar{m}^{-1}\left(\bar{x}, w, \frac{u}{\bar{u}}\alpha\right)\right).$$

N.4 (Transformation model) Suppose $d_x \geq 2$ and partition $X = (X_1, X_2)$ with $\mathcal{X}_1 \subseteq \mathbb{R}^{d_x - 1}, \mathcal{X}_2 \subseteq \mathbb{R}$. Assume that there exists a (possibly unknown) function

[33] I have presented these results as the minimum needed for identification when $U \not\perp\!\!\!\perp W$. When $U \perp\!\!\!\perp W$, the results can be strengthened to hold only at specific $(\bar{x}, \bar{w})$ pairs.

$\mu(x_1, w, t)$ that is strictly increasing in t, such that $m(\bar{x}_1, x_2, w, u) = \mu(\bar{x}_1, w, u - x_2)$ for all x_2, u and some known $\bar{x}_1$. Also assume that $\mu(\bar{x}_1, w, \alpha) = \bar{y}$ for some known $\alpha \in \mathbb{R}, \bar{y} \in \mathbb{R}$ and all w. Then

$$m(x, w, u) = \bar{m}(x, w, \bar{m}^{-1}(\bar{x}_1, u - \alpha, w, \bar{y})).$$

The intuitive reason that some normalization is needed is that since F_U is unknown, the scale (or magnitude) of U is unknown. Matzkin (2003) presents this formally, but a simple example is a convincing heuristic: $m_1(x, u)$ when $U \sim \text{Unif}[0, 1]$ is observationally equivalent to $m_2(x, u) = m_1(x, u/2)$ when $U \sim \text{Unif}[0, 2]$ even though $m_1 \neq m_2$. Normalization N.1 solves this by specifying an *a priori* distribution of U, whereas N.2 solves this by tying the scale of U, which is unobserved, to the scale of Y, which is observed.

The homogeneity of degree one normalization in N.3 achieves the same result by fixing the scale of m at a known point $(\bar{x}, \bar{w}, \bar{u})$ and using a shape restriction to ensure that this is sufficient to fix the scale of m over the entire support of U, whatever it may be. This can be a particularly attractive normalization if m represents a profit or cost function, as economic theory suggests several circumstances under which these functions should be homogeneous of degree one. For example, if m is a cost function of output quantity X and input prices (W, U), one of which is unobserved, then m will be homogeneous of degree one in (W, U) if the firm minimizes costs and takes input prices as given.

Normalization N.4 is a generalization of an additive transformation model that imposes additive separability for one component of X. Matzkin (2003) shows how this specification can be derived from a consumer demand problem.

Proof of Proposition B.1. All of these derivations follow the same strategy. First, I show that the normalization implies that the distribution of U is identified from $\bar{m}$. Then using $\bar{m}(x, w, s) = m(x, w, Q_{U|W}(s \mid w))$, or the equivalent $m(x, w, u) = \bar{m}(x, w, F_{U|W}(u \mid w))$, I constructively identify m.

For N.2,

$$\bar{m}(\bar{x}, w, F_{U|W}(u \mid w)) = m(\bar{x}, w, u) = u,$$

hence $F_{U|W}(u \mid w) = \bar{m}^{-1}(\bar{x}, w, u)$, with $\bar{m}^{-1}$ the inverse of $\bar{m}$ in its last argument, and

$$m(x, w, u) = \bar{m}(x, w, F_{U|W}(u \mid w)) = \bar{m}(x, w, \bar{m}^{-1}(\bar{x}, w, u)).$$

When maintaining N.3,

$$\bar{m}\left(\bar{x}, w, F_{U|W}(u \mid w)\right) = \bar{m}\left(\bar{x}, w, F_{U|W}\left(\frac{u}{\bar{u}}\bar{u} \,\middle|\, w\right)\right) = m\left(\bar{x}, w, \frac{u}{\bar{u}}\bar{u}\right) = (u/\bar{u})\alpha,$$

and so

$$F_{U|W}(u \mid w) = \bar{m}^{-1}\left(\bar{x}, w, \frac{u}{\bar{u}}\alpha\right) \quad \text{and} \quad m(x, w, u) = \bar{m}\left(x, w, \bar{m}^{-1}\left(\bar{x}, w, \frac{u}{\bar{u}}\alpha\right)\right).$$

Under N.4,

$$\begin{aligned}\bar{m}(\bar{x}_1, Q_{U|W}(s \mid w) - \alpha, w, s) &= m\left(\bar{x}_1, Q_{U|W}(s \mid w) - \alpha, w, Q_{U|W}(s \mid w)\right) \\ &= \mu(\bar{x}_1, w, \alpha) = \bar{y}.\end{aligned}$$

Hence $\bar{m}(\bar{x}_1, u - \alpha, w, F_{U|W}(u \mid w)) = \bar{y}$, so $F_{U|W}(u \mid w) = \bar{m}^{-1}(\bar{x}_1, u - \alpha, w, \bar{y})$ and

$$m(x, w, u) = \bar{m}(x, w, \bar{m}^{-1}(\bar{x}_1, u - \alpha, w, \bar{y})).$$

Q.E.D.

C Proofs for Sections 2-4

For Propositions 1-4 and Theorems 2-3 it may not be clear how to extend the results for covariates, so, at the cost of some elegance, I explicitly prove the general versions given in Appendix A and Section 4. For Theorem 1 and Proposition 5, I leave the conditioning on covariates implicit.

To compress notation I refer to an event like $[X = x, W = w]$ merely as $[x, w]$, but I break this convention when it would be ambiguous, or when it is desirable for emphasis.

Proposition C.1. *Suppose X is a d_x-dimensional random vector, U is a random variable and Z is a random element such that $X|Z = z$ is continuously distributed for any $z \in \mathcal{Z}$. Then for any u and any $(x, z) \in \mathcal{X}\mathcal{Z}$,*

$$F_{U|XZ}(u \mid x, z) = C_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), F_{U|Z}(u \mid z); z\right)\sigma(x|z),$$

where $C(\cdot, \cdot; z)$ is the copula of the random vector $(X, U)|Z = z$, $C_{\mathbf{X}}$ is the d_x^{th}-order partial derivative of C with respect to each of the first d_x arguments, $\mathbf{F}_{X|Z}(x \mid z) =$

$(F_{X_1|Z}(x_1 \mid z), \ldots, F_{X_{d_x}|Z}(x_{d_x} \mid z))'$ *is the vector of marginal distribution functions and*

$$\sigma(x|z) = \frac{\prod_{k=1}^{d_x} f_{X_k|Z}(x_k \mid z)}{f_{X|Z}(x \mid z)},$$

where $f_{X_k|Z}$ *and* $f_{X|Z}$ *are densities. Moreover, for any single* k*,*

$$F_{U|X_kZ}(u \mid x_k, z) = C_k\left(\mathbf{F}^{\mathbf{1}}_{X_k|Z}(x_k \mid z), F_{U|Z}(u \mid z); z\right),$$

where C_k *is the partial derivative of* C *with respect to its* k^{th} *argument and* $\mathbf{F}^{\mathbf{1}}_{X_k|Z}(x_k|z) = \left(1, \ldots, F_{X_k|Z}(x_k \mid z), \ldots, 1\right)' \in \{1\}^{k-1} \times [0,1] \times \{1\}^{d_x-k}$ *is* 1 *in every component but the* k^{th}*.*

Note that analogous statements also hold for the unconditional copula of (X, U)*. To see this, simply condition on the entire underlying sample space, rather than the event* $[Z = z]$*.*

Proof. It follows from the definition of a conditional distribution function that for any u and any $(x, z) \in \mathcal{X}\mathcal{Z}$,

$$\begin{aligned} F_{U|XZ}(u \mid x, z) &= \frac{\nabla^{d_x}_{x_1,\ldots,x_{d_x}} F_{XU|Z}(x, u \mid z)}{f_{X|Z}(x \mid z)} \\ &= C_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), F_{U|Z}(u \mid z); z\right) \frac{\prod_{k=1}^{d_x} f_{X_k|Z}(x_k \mid z)}{f_{X|Z}(x \mid z)}, \end{aligned}$$

where the second equality is from the conditional form of Sklar's Theorem (footnote 5). The result for any single k then follows by noting that for the vector $x_k^\infty = (+\infty, \ldots, x_k, \ldots, +\infty)' \in \{+\infty\}^{k-1} \times \bar{\mathbb{R}} \times \{+\infty\}^{d_x-k}$,

$$\begin{aligned} F_{U|X_kZ}(u \mid x_k, z) &= \frac{\nabla_{x_k} F_{X_kU|Z}(x_k, u \mid z)}{f_{X_k|Z}(x_k \mid z)} \\ &= \frac{\nabla_{x_k} F_{XU|Z}(x_k^\infty, u \mid z)}{f_{X_k|Z}(x_k \mid z)} = C_k\left(\mathbf{F}^{\mathbf{1}}_{X_k|Z}(x_k \mid z), F_{U|Z}(u \mid z); z\right). \end{aligned}$$

Q.E.D.

Proof of Proposition 1′. Let $r, u \in [0,1]$, $(w, z) \in \mathcal{W}\mathcal{Z}$. Let $\mathbf{r}^{\mathbf{1}}_k = (1, \ldots, r, \ldots, 1)' \in \{1\}^{k-1} \times [0,1] \times \{1\}^{d_x-k}$ be a vector of ones with r in the k^{th} position. Also, recall the similar definition $\mathbf{F}^{\mathbf{1}}_{X_k|WZ}$ from Proposition C.1. Note that because of M.2′, $(w, z, r) \in$

$\text{supp}(W, Z, R_k) = \mathcal{WZ} \times [0,1]$ for every k. Then

$$\begin{aligned}
\mathbb{P}\left[U \leq u \mid w, z, R_k = r\right] &= \mathbb{P}\left[U \leq u \,\middle|\, w, z, F_{X_k|WZ}(X_k \mid W, Z) = r\right] \\
&= \mathbb{P}\left[U \leq u \,\middle|\, w, z, X_k = Q_{X_k|WZ}(r \mid w, z)\right] \\
&= C_k\left(\mathbf{F}^{\mathbf{1}}_{X_k|WZ}(Q_{X_k|WZ}(r \mid w, z) \mid w, z), F_{U|WZ}(u \mid w, z); (w, z)\right) \\
&= C_k\left(\mathbf{r}^{\mathbf{1}}_k, F_{U|W}(u \mid w); (w, z)\right),
\end{aligned}$$

where the second equality follows from M.2′, the third equality follows from M.2′ with Proposition C.1 and the fourth equality is due to M.3′ and the definition of $Q_{X_k|WZ}$. Hence for every k, every supported (w, z, r) and every u, $\mathbb{P}\left[U \leq u \mid w, z, R_k = r\right]$ does not depend on z, that is $U \perp\!\!\!\perp Z | W, R_k$, if and only if $C_k(\mathbf{r}^{\mathbf{1}}_k, F_{U|W}(u \mid w); (w, z)) = C_k(\mathbf{r}^{\mathbf{1}}_k, F_{U|W}(u \mid w); w)$, which occurs if and only if $C_k\left(\cdot, \cdot; (w, z)\right) = C_k\left(\cdot, \cdot; w\right)$ for all $(w, z) \in \mathcal{WZ}$.

Finally, since any copula satisfies $C\left(r, u\right) = 0$ whenever any component of r is 0, it follows from the fundamental theorem of calculus that the previous statement is equivalent to $C(\cdot, \cdot; (w, z)) = C(\cdot, \cdot; w)$. Formally, for any $r \in [0,1]^{d_x}, u \in [0,1]$ and any $(w, z) \in \mathcal{WZ}$,

$$\begin{aligned}
C\left(r, u; (w, z)\right) &= \int_0^{r_{d_x}} \cdots \int_0^{r_1} C_{\mathbf{X}}(s, u; (w, z)) ds_1 \cdots ds_{d_x} \\
&= \int_0^{r_{d_x}} \cdots \int_0^{r_1} C_{\mathbf{X}}(s, u; w) ds_1 \cdots ds_{d_x} = C\left(r, u; w\right),
\end{aligned}$$

which is CCI′. Q.E.D.

Proof of Proposition 2′. Suppose that $(x_k, w, r) \in \text{supp}(X_k, W, R_k)$ and $u \in [0,1]$. Define $\mathbf{r}^{\mathbf{1}}_k$ as in the proof of Proposition 1′ and $\mathbf{F}^{\mathbf{1}}_{X_k|WZ}$ as in Proposition C.1. Then for every k,

$$\begin{aligned}
& \mathbb{P}\left(U \leq u \mid w, R_k = r, x_k\right) \\
= \;& \mathbb{E}\left[\mathbb{P}\left(U \leq u \mid w, r, x_k, Z\right) \mid w, r, x_k\right] \\
= \;& \mathbb{E}\left[C_k\left(\mathbf{F}^{\mathbf{1}}_{X_k|WZ}(Q_{X_k|WZ}(r \mid w, Z) \mid w, Z), F_{U|W}(u \mid w); (w, Z)\right) \,\middle|\, w, r, x_k\right] \\
= \;& \mathbb{E}\left[C_k\left(\mathbf{r}^{\mathbf{1}}_k, F_{U|W}(u \mid w); (w, Z)\right) \mid w, r, x_k\right] \\
= \;& C_k(\mathbf{r}^{\mathbf{1}}_k, F_{U|W}(u \mid w); w), \qquad (17)
\end{aligned}$$

where the second equality uses M.2′ with Proposition C.1 and M.3′, the third equality is from the definition of $Q_{X_k|WZ}$ and the fourth equality is due to CCI′. Hence

$\mathbb{P}[U \leq u \mid w, R_k = r, x_k]$ does not depend on x_k for any $(x_k, w, r) \in \text{supp}(X_k, W, R_k)$, i.e. $U \perp\!\!\!\perp X_k | W, R_k$.

A trivial example in which the converse does not hold is when $d_x = 1$ and $X \perp\!\!\!\perp Z$. In this case $R = Q_X(X)$ and the event $[X = x, R = r]$, for $(x, r) \in \text{supp}(X, R)$, is equivalent to the event $[R = r]$ by M.2. Thus $\mathbb{P}[U \leq u \mid X = x, R = r] = \mathbb{P}[U \leq u \mid R = r]$ and $U \perp\!\!\!\perp X | R$, regardless of whether or not CCI holds. This counterexample is not interesting because Assumption R does not hold if $X \perp\!\!\!\perp Z$.

The following is a non-trivial counterexample. Suppose that $\mathcal{Z} = \{A_0, A_1, B_0, B_1\}$ and that Z assumes these values with equal probability, independently of U. Let $X = \mathbb{1}[Z \in \{A_1, B_1\}] + V$, where $V \sim \text{Unif}[0, 1]$ and $V \perp\!\!\!\perp Z$. Thus $R = V$ is the conditional ranking of X and the support of X given $R = r$ is $\{r, 1 + r\}$ with these two points of support corresponding to whether $Z \in \{A_0, B_0\}$ or $Z \in \{A_1, B_1\}$. Assumptions M.2 and M.3 are satisfied in this setup.

Now suppose that $C(\cdot, \cdot; A_0) = C(\cdot, \cdot; A_1)$ and $C(\cdot, \cdot; B_0) = C(\cdot, \cdot; B_1)$, but that $C(r, u; A_0) \neq C(r, u; B_0)$ for some $(r, u) \in (0, 1)^2$ so that CCI does not hold. Nevertheless, it is still the case that $U \perp\!\!\!\perp X | R$. To see this, note that since $R = V \perp\!\!\!\perp Z$ and $[R = r, X = r] = [R = r, Z \in \{A_0, B_0\}] = [V = r, Z \in \{A_0, B_0\}]$, it follows that $\mathbb{P}[Z = A_0 \mid R = r, X = r] = \mathbb{P}[Z = B_0 \mid R = r, X = r] = 1/2$. For the same reason, $\mathbb{P}[Z = A_1 \mid R = r, X = 1 + r] = \mathbb{P}[Z = B_1 \mid R = r, X = 1 + r] = 1/2$. Thus

$$\begin{aligned}
\mathbb{E}[C_1(r, u; Z) \mid R = r, X = r] &= \frac{1}{2} C_1(r, u; A_0) + \frac{1}{2} C_1(r, u; B_0) \\
&= \frac{1}{2} C_1(r, u; A_1) + \frac{1}{2} C_1(r, u; B_1) \\
&= \mathbb{E}[C_1(r, u; Z) \mid R = r, X = 1 + r].
\end{aligned}$$

Since given $R = r$, the support of X is $\{r, 1 + r\}$, this shows that $U \perp\!\!\!\perp X | R$ (see (17)), despite CCI not holding.

It is interesting to consider how this counterexample fails the sufficient conditions in Theorem 2. Assumption S.2 is satisfied and both $U \perp\!\!\!\perp Z$ and $V \perp\!\!\!\perp Z$ by assumption. However, $(U, V) \not\perp\!\!\!\perp Z$, so S.1 does not hold. It suffices to note that $U \not\perp\!\!\!\perp Z | V$, which would not be the case if $(U, V) \perp\!\!\!\perp Z$. This is because for some $(r, u) \in (0, 1)^2$,

$$\mathbb{P}[U \leq u \mid V = r, Z = A_0] = C_1(r, u; A_0) \neq C_1(r, u; B_0) = \mathbb{P}[U \leq u \mid V = r, Z = B_0].$$

Thus CCI is closely related to the basic observation that $(U, V) \perp\!\!\!\perp Z$ implies but is not implied by $U \perp\!\!\!\perp Z$ and $V \perp\!\!\!\perp Z$. In fact, as can be seen from the conditional form of Sklar's Theorem, CCI is precisely what is needed to make this statement an equivalence.

Q.E.D.

Proof of Proposition 3′. From Proposition C.1, which requires M.2′,

$$F_{U|XW}(u \mid x, w) = C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), F_{U|W}(u \mid w); w\right)\sigma(x|w)$$

for any u, any $(x, w) \in \mathcal{X}\mathcal{W}$. By M.1′,

$$\begin{aligned} F_{U|XW}(u \mid x, w) &= \mathbb{P}[U \leq u \mid x, w] \\ &= \mathbb{P}[m(x, w, U) \leq m(x, w, u) \mid x, w] \\ &= \mathbb{P}[Y \leq m(x, w, u) \mid x, w] = F_{Y|XW}(m(x, w, u) \mid x, w). \end{aligned}$$

Hence

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), F_{U|W}(u \mid w); w\right) = F_{Y|XW}(m(x, w, u) \mid x, w)/\sigma(x|w).$$

By taking $u = Q_{U|W}(s \mid w)$ for any $s \in (0, 1)$, one obtains

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), s; w\right) = F_{Y|XW}(\bar{m}(x, w, s) \mid x, w)/\sigma(x|w).$$

M.1′ ensures that $F_{Y|XW}(\cdot \mid x, w)$ is invertible with inverse $Q_{Y|XW}(\cdot \mid x, w)$, from which it follows that

$$\bar{m}(x, w, s) = Q_{Y|XW}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|W}(x \mid w), s\right)\sigma(x|w) \,\middle|\, x, w\right).$$

Q.E.D.

Proof of Proposition 4′. For any u, any $(x, w, z) \in \mathcal{X}\mathcal{W}\mathcal{Z}$,

$$\begin{aligned} F_{U|XWZ}(u \mid x, w, z) &= C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}(x \mid w, z), F_{U|WZ}(u \mid w, z); (w, z)\right)\sigma(x|w, z) \\ &= C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}(x \mid w, z), F_{U|W}(u \mid w); (w, z)\right)\sigma(x|w, z), \\ &= C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}(x \mid w, z), F_{U|W}(u \mid w); w\right)\sigma(x|w, z), \end{aligned}$$

where the first equality follows from M.2′ and Proposition C.1, the second equality follows from M.3′ and the third equality follows from CCI′. The rest of the proof is

now similar to that for Proposition 3. By M.1′,

$$\begin{aligned} F_{U|XWZ}\left(u \mid x,w,z\right) &= \mathbb{P}\left[U \leq u \mid x,w,z\right] \\ &= \mathbb{P}\left[m(x,w,U) \leq m(x,w,u) \mid x,w,z\right] \\ &= \mathbb{P}\left[Y \leq m(x,w,u) \mid x,w,z\right] = F_{Y|XWZ}\left(m(x,w,u) \mid x,w,z\right). \end{aligned}$$

Hence

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}\left(x \mid w,z\right), F_{U|W}(u \mid w); w\right) = F_{Y|XWZ}\left(m(x,w,u) \mid x,w,z\right)/\sigma(x|w,z),$$

and so by taking $u = Q_{U|W}(s \mid w)$,

$$C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}\left(x \mid w,z\right), s; w\right) = F_{Y|XWZ}\left(\bar{m}(x,w,s) \mid x,w,z\right)/\sigma(x|w,z).$$

M.1′ ensures that $F_{Y|XWZ}\left(\cdot \mid x,w,z\right)$ is invertible with inverse $Q_{Y|XWZ}\left(\cdot \mid x,w,z\right)$, from which it follows that

$$\bar{m}(x,w,s) = Q_{Y|XWZ}\left(C_{\mathbf{X}}\left(\mathbf{F}_{X|WZ}\left(x \mid w,z\right), s; w\right)\sigma(x|w,z) \,\middle|\, x,w,z\right).$$

Q.E.D.

Proof of Theorem 1. In what follows, I continue to assume that $\mathbb{P}\,\mathcal{X}^D = 1$ and to ignore the distinction between X^D and X. See the discussion following the statement of this theorem and Proposition 5 for the reasoning and justification behind this approach. To keep the notation reasonable, I leave conditioning on covariates implicit.

I require the following lemma, which describes an implication of local dependence for continuous random variables.

Lemma C.1. *Suppose $X|Z = z$ is continuously distributed for all $z \in \mathcal{Z}$. If $X \not\perp\!\!\!\perp Z$ at x^* then there exist $z_k^*, k = 1, \ldots, d_x$ and a $\delta > 0$ such that $\left|F_{X_k}(x_k) - F_{X_k|Z}(x_k \mid z_k^*)\right| > 0$ for all k and all $x \in B_\delta(x^*)$.*

Proof of Lemma C.1. Suppose $X \not\perp\!\!\!\perp Z$ at x^*, i.e. $X_k \not\perp\!\!\!\perp Z$ at x_k^* for every k. Thus $\mathbb{P}\,\mathcal{Z}_k(x_k^*) > 0$ for every k, so take any $z_k^* \in \mathcal{Z}_k(x_k^*)$ and let $\left|F_{X_k}(x_k^*) - F_{X_k|Z}(x_k^* \mid z_k^*)\right| = \epsilon_k^* > 0$. Since $F_{X_k}(\cdot)$ and $F_{X_k|Z}(\cdot \mid z_k^*)$ are continuous functions, there exists a $\delta_k > 0$ such that $|F_{X_k}(x_k^*) - F_{X_k}(x_k)| < \epsilon_k^*/2$ and $\left|F_{X_k|Z}(x_k^* \mid z_k^*) - F_{X_k|Z}(x_k \mid z_k^*)\right| < \epsilon_k^*/2$ for

any $x_k \in B_{\delta_k}(x_k^*)$. Let $\delta = \min_k \delta_k$. Then for any $x \in B_\delta(x^*)$ and every k,

$$\begin{aligned}\epsilon_k^* = \left|F_{X_k}(x_k^*) - F_{X_k|Z}(x_k^* \mid z_k^*)\right| &\leq |F_{X_k}(x_k^*) - F_{X_k}(x_k)| + \left|F_{X_k}(x_k) - F_{X_k|Z}(x_k \mid z_k^*)\right| \\ &\quad + \left|F_{X_k|Z}(x_k \mid z_k^*) - F_{X_k|Z}(x_k^* \mid z_k^*)\right| \\ &< \epsilon_k^* + \left|F_{X_k}(x_k) - F_{X_k|Z}(x_k \mid z_k^*)\right|\end{aligned}$$

and so $\left|F_{X_k}(x_k) - F_{X_k|Z}(x_k \mid z_k^*)\right| > 0$. Q.E.D.

The proof strategy for Theorem 1 is the same as in the special case presented in the main text. Recall that $T(C') = 0$ implies that for every $s \in (0,1)$ and a.e. $(x,z) \in \mathcal{X}\mathcal{Z}$,

$$C_{\mathbf{X}}^{-1}\left(\mathbf{F}_X(x), C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\right) = C_{\mathbf{X}}^{-1}\left(\mathbf{F}_{X|Z}(x \mid z), C'_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z), s\right)\right), \tag{11}$$

which is defined as $I(x,s)$. First, I show that $I(x,s)$ is not a function of x for $x \in \mathcal{X}_D$ and all $s \in (0,1)$. Second, I show that this can only be the case if $I(x,s) = I(s) = s$ for all s. Identification of $\bar{m}$ is then achieved via (11) and Proposition 3.

Step 1: Take any $x^* \in \mathcal{X}$ such that $X \not\perp\!\!\!\perp Z$ at x^*, i.e. any $x^* \in \mathcal{X}_D$. By M.2 and Lemma C.1, there exist $\delta > 0$ and z_k^*, $k = 1, \dots d_x$, such that for any $x^0 \in B_\delta(x^*)$, $\left|F_{X_k}(x_k^0) - F_{X_k|Z}(x_k^0 \mid z_k^*)\right| > 0$ for every k. I will show that $I(x^0, s) = I(\xi^0, s)$ for any $\xi^0 = (x_1^0, \dots, x_j^0 + \gamma, \dots, x_{d_x}^0)$ where $\gamma > 0$ is a small number such that $\xi^0 \in B_\delta(x^*)$. This shows that $I(x,s)$ does not depend on x_j over $B_\delta(x^*)$. Repeating the argument for every $j = 1, \dots, d_x$ establishes that $I(x,s)$ does not depend on x over $B_\delta(x^*)$ and hence that $I(x,s)$ does not depend on x over $\mathcal{X}_D$. Since the proof is symmetric in j, I take $j = 1$ without loss of generality. I also assume for concreteness that $F_{X_1}(x_1^0) - F_{X_1|Z}(x_1^0 \mid z_1^*) > 0$; the changes required for the opposite case will be clear. By M.2, γ can be taken small enough so that $F_{X_1}(\xi_1^0) - F_{X_1|Z}(\xi_1^0 \mid z_1^*) > 0$ as well.

Consider the mappings

$$q_k : \mathcal{X}_k \to \mathcal{X}_k : q(x_k) = Q_{X_k}(F_{X_k|Z}(x_k \mid z_1^*))$$

for $k = 1, \dots, d_x$, $q(x) = (q_1(x_1), \dots, q_{d_x}(x_{d_x}))' \in \bar{\mathbb{R}}^{d_x}$, and the associated sequence $\{x^n\}$ formed as $x^n = q(x^{n-1})$ for $n \geq 1$.[34] For each k, $\{x_k^n\}$ is a weakly monotone sequence that is either increasing or decreasing depending on the sign of $\iota_k = F_{X_k}(x_k^0) - F_{X_k|Z}(x_k^0 \mid z_1^*)$. For example, if $\iota_k \geq 0$ then

$$F_{X_k}(x_k^1) = F_{X_k}(q_k(x_k^0)) = F_{X_k|Z}(x_k^0 \mid z_1^*) \leq F_{X_k}(x_k^0),$$

[34]Note that this definition of q is different than the related definition $q(\cdot|z)$, which is used in Sections 3 and 5. This should cause no confusion.

which shows that $x_k^1 \leq x_k^0$ since F_{X_k} is strictly increasing by M.2. Generally, if $x_k^n \leq x_k^{n-1}$ then

$$F_{X_k}(x_k^{n+1}) = F_{X_k}(q(x_k^n)) = F_{X_k|Z}(x_k^n \mid z_1^*) \leq F_{X_k|Z}(x_k^{n-1} \mid z_1^*) = F_{X_k}(x_k^n)$$

and so $x_k^{n+1} \leq x_k^n$. It follows by induction that $\{x_k^n\}$ is weakly decreasing if $\iota_k \geq 0$. If $\iota_k \leq 0$ then $\{x_k^n\}$ is a weakly increasing sequence. In either case, $\{x_k^n\}$ has a limiting point $x_k^L \in \bar{\mathbb{R}}$, which may be $\pm\infty$. Hence $\lim_n x^n = x^L = (x_1^L, \ldots, x_{d_x}^L)$ exists. Since $\iota_1 > 0$ (by innocuous assumption), $\{x_1^n\}$ is strictly decreasing and $x_1^L < x_1^0$.

From M.2 it follows that $x_1^L \in \bar{\mathcal{X}}_1 = \left\{x_1 < x_1^0 : F_{X_1}(x_1) = F_{X_1|Z}(x_1 \mid z_1^*)\right\}$ since

$$F_{X_1}(x_1^L) = \lim_n F_{X_1}(x_1^{n+1}) = \lim_n F_{X_1|Z}(x_1^n \mid z_1^*) = F_{X_1|Z}(x_1^L \mid z_1^*).$$

By the definition of δ, $x^L \notin B_\delta(x^*)$. Moreover, $x_1^L = \sup \bar{\mathcal{X}}_1$. For if there did exist an $\bar{x}_1 \in \bar{\mathcal{X}}_1$ such that $\bar{x}_1 > x_1^L$ then because $\{x_1^n\}$ is strictly decreasing there would exist an N such that $x_1^N \geq \bar{x}_1 > x_1^{N+1} > x_1^L$ and this would imply that

$$F_{X_1|Z}(x_1^N \mid z_1^*) = F_{X_1}(x_1^{N+1}) < F_{X_1}(\bar{x}_1) = F_{X_1|Z}(\bar{x}_1 \mid z_1^*),$$

which is a contradiction since $F_{X_1|Z}(\cdot \mid z_1^*)$ is strictly increasing. Thus $x_1^L = \sup \bar{\mathcal{X}}_1$, as claimed.

Now consider the analogous sequence $\{\xi^n\}$ formed as $\xi^n = q(\xi^{n-1})$ for $n \geq 1$. Notice that $\xi_k^n = x_k^n$ for $k \neq 1$, so these components of the sequence trivially have the same limiting point. The previous analysis applied to $\{\xi_1^n\}$ shows that it is also a strictly decreasing sequence with limiting point $\xi_1^L = \sup \bar{\mathcal{X}}_1$. Therefore $\xi_1^L = x_1^L$ and hence $\xi^L = x^L$.

Finally, observe that $I(x, s) = I(q(x), s)$ since

$$\begin{aligned} I(x,s) &= C_{\mathbf{X}}^{-1}\left(\mathbf{F}_X(x), C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\right) \\ &= C_{\mathbf{X}}^{-1}\left(\mathbf{F}_{X|Z}(x \mid z_1^*), C'_{\mathbf{X}}\left(\mathbf{F}_{X|Z}(x \mid z_1^*), s\right)\right) \\ &= C_{\mathbf{X}}^{-1}\left(\mathbf{F}_X(q(x)), C'_{\mathbf{X}}\left(\mathbf{F}_X(q(x)), s\right)\right) = I(q(x), s). \end{aligned}$$

Hence $I(x^0, s) = I(x^n, s)$ and $I(\xi^0, s) = I(\xi^n, s)$ for all n and all s. Since $I(x, s)$ is a continuous function of x by M.1 and M.2, $I(x^0, s) = \lim_n I(x^n, s) = I(x^L, s) = I(\xi^L, s) = \lim_n I(\xi^n, s) = I(\xi^0, s)$. Thus $I(x^0, s) = I(\xi^0, s)$ for all s, as claimed.

Step 2: Return once more to

$$I(x,s) = C_{\mathbf{X}}^{-1}\left(\mathbf{F}_X(x), C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)\right), \tag{11}$$

which can be written equivalently as

$$C_{\mathbf{X}}\left(\mathbf{F}_X(x), I(x,s)\right) = C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right). \tag{18}$$

Since by definition a copula has unit-uniform marginal distributions, it follows that for any copula $C_{\mathbf{X}}$ and any $t \in (0,1)$,

$$\int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{d_x}} C_{\mathbf{X}}\left(\mathbf{F}_X(x), t\right) \left[\prod_{k=1}^{d_x} f_{X_k}(x_k)\right] dx_{d_x} \cdots dx_1 = t. \tag{19}$$

Since $I(x,s) = I(s)$ for all $x \in \mathcal{X}_D$ and $\mathbb{P}\,\mathcal{X}_D = 1$, weighting (18) by $\prod_{k=1}^{d_x} f_{X_k}(x_k)$, integrating both sides over $\mathcal{X}_D$ and using (19) yields $I(s) = s$. Thus from (18), $C_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right) = C'_{\mathbf{X}}\left(\mathbf{F}_X(x), s\right)$, for all s. This shows that $C_{\mathbf{X}}(F_X(x), s)$ is identified a.e.-F_X. By Proposition 3, this shows that $\bar{m}(x,s)$ is identified a.e.-F_X. Q.E.D.

Proof of Proposition 5. Define $R_k^D = F_{X_k^D|Z}(X_k^D \mid Z)$. By Proposition 1′, CCI holds for (X, U, Z) if and only if $U \perp\!\!\!\perp Z | R_k$ for all k. I will show that $U \perp\!\!\!\perp Z | R_k$ implies $U \perp\!\!\!\perp Z | R_k^D$ and hence that CCI holds for (X^D, U, Z) as well. Since the argument does not depend on the component k, I drop the subscript.

I first derive an expression for $F_{X^D|Z}(x \mid z)$ in terms of $F_{X|Z}(x \mid z)$ for any $z \in \mathcal{Z}$. Note that since $X \perp\!\!\!\perp Z$ at x for all $x \notin \mathcal{X}^D$, $\mathbb{P}\left[X \notin \mathcal{X}^D \,\middle|\, Z = z\right] = \mathbb{P}\left[X \notin \mathcal{X}^D\right]$ for all $z \in Z$. Hence $p^D = \mathbb{P}\left[X \in \mathcal{X}^D \,\middle|\, Z = z\right] = 1 - \mathbb{P}\left[X \notin \mathcal{X}^D\right]$ does not depend on z either. Define the set $\mathcal{X}^L(x) = \{x' \notin \mathcal{X}^D : F_X(x') \leq F_X(x)\}$ and $x^L(x) = \sup \mathcal{X}^L(x)$. Let $p^L(x) = F_X(x^L(x))$. Since $x^L(x)$ is in the closure of $\mathcal{X} \setminus \mathcal{X}^D$, M.2 ensures that $p^L(x) = F_{X|Z}(x^L(x) \mid z)$ for every z. Also, note that

$$p^{DL}(x) = \mathbb{P}\left[X \leq x^L(x), X \in \mathcal{X}^D \,\middle|\, z\right] = F_X(x^L(x)) - \mathbb{P}\left[X \leq x^L(x), X \notin \mathcal{X}^D \,\middle|\, z\right]$$

does not depend on z either since $X \perp\!\!\!\perp Z$ at x for all $x \notin \mathcal{X}^D$. These quantities relate $F_{X^D|Z}(x \mid z)$ to $F_{X|Z}(x \mid z)$ through the following argument.

$$\begin{aligned}
F_{X^D|Z}(x \mid z) &= \mathbb{P}\left[X^D \leq x \,\middle|\, z\right] \\
&= \mathbb{P}\left[X \leq x \,\middle|\, z, X \in \mathcal{X}^D\right] \\
&= \mathbb{P}\left[X \leq x, X \in \mathcal{X}^D \,\middle|\, z\right] / p^D \\
&= \left(\mathbb{P}\left[X \in [x^L(x), x], X \in \mathcal{X}^D \,\middle|\, z\right] + \mathbb{P}\left[X \leq x^L(x), X \in \mathcal{X}^D \,\middle|\, z\right]\right) / p^D \\
&= \left(\mathbb{P}\left[X \in [x^L(x), x] \,\middle|\, z\right] + \mathbb{P}\left[X \leq x^L(x), X \in \mathcal{X}^D \,\middle|\, z\right]\right) / p^D \\
&= \left(F_{X|Z}(x \mid z) - p^L(x) + p^{DL}(x)\right) / p^D,
\end{aligned} \tag{20}$$

where the second to last equality is because $[x^L(x), x] \subseteq \mathcal{X}^D$ by construction.

To use (20) to show that $U \perp\!\!\!\perp Z | R^D$, I will first establish that for any given r, $F_{X|Z}(Q_{X^D|Z}(r \mid z) \mid z)$ does not depend on z. From (20),

$$F_{X|Z}(Q_{X^D|Z}(r \mid z) \mid z) = r/p^D + p^L(Q_{X^D|Z}(r \mid z)) + p^{DL}(Q_{X^D|Z}(r \mid z)). \tag{21}$$

Hence, given the definitions of p^L and p^{DL}, it suffices to show that $x^L(Q_{X^D|Z}(r \mid z))$ does not depend on z. Note first that $X \perp\!\!\!\perp Z$ at x' implies that $X^D \perp\!\!\!\perp Z$ at x' as well, since in this case $x' = x^L(x')$ and hence by (20), $F_{X^D|Z}(x' \mid z) = p^{DL}(x')/p^D$, which does not depend on z. By the monotonicity of distribution functions, for any r and z, $x' \leq Q_{X^D|Z}(r \mid z)$ if and only if $F_{X^D|Z}(x' \mid z) \leq F_{X^D|Z}(Q_{X^D|Z}(r \mid z) \mid z) = r$. Since $X^D \perp\!\!\!\perp Z$ at x' implies that $F_{X^D|Z}(x' \mid z) = F_{X^D|Z}(x' \mid z')$, this occurs if and only if $x' \leq Q_{X^D|Z}(r \mid z')$. Hence for any z, z', one has $x' \in \mathcal{X}^L(Q_{X^D|Z}(r \mid z))$ if and only if $x' \in \mathcal{X}^L(Q_{X^D|Z}(r \mid z'))$. It follows that $x^L(Q_{X^D|Z}(r \mid z)) = \sup \mathcal{X}^L(Q_{X^D|Z}(r \mid z)) = \sup \mathcal{X}^L(Q_{X^D|Z}(r \mid z')) = x^L(Q_{X^D|Z}(r \mid z'))$ for any z, z' as well, i.e. $x^L(Q_{X^D|Z}(r \mid z))$ does not depend on z.

Finally, consider the event $[Z = z, R^D = r]$ for any $r \in [0,1]$ and any $z \in Z$. Notice that while $Q_{X^D|Z}(r|z)$ in general depends on z for any given r, the preceding argument shows that (21) does not depend on z, given r. Since $X^D = Q_{X^D|Z}(r \mid z)$ if and only if $X = Q_{X^D|Z}(r \mid z)$, it follows that

$$\begin{aligned}
\mathbb{P}\left[U \leq u \,\middle|\, z, R^D = r\right] &= \mathbb{P}\left[U \leq u \,\middle|\, z, F_{X^D|Z}(X^D \mid z) = r\right] \\
&= \mathbb{P}\left[U \leq u \,\middle|\, z, X^D = Q_{X^D|Z}(r \mid z)\right] \\
&= \mathbb{P}\left[U \leq u \,\middle|\, z, X = Q_{X^D|Z}(r \mid z)\right] \\
&= \mathbb{P}\left[U \leq u \,\middle|\, z, F_{X|Z}(X \mid z) = F_{X|Z}(Q_{X^D|Z}(r \mid z) \mid z)\right] \\
&= \mathbb{P}\left[U \leq u \mid z, R = (21)\right]
\end{aligned}$$

does not depend on z, since $U \perp\!\!\!\perp Z | R$. This shows that $U \perp\!\!\!\perp Z | R^D$ and hence that CCI holds for (X^D, U, Z) as well. Q.E.D.

Proof of Theorem 2. By Proposition 1′, it suffices to verify that $U \perp\!\!\!\perp Z | R_k, W$ for every k. Consider any $(w, z), (w, z') \in \mathcal{WZ}$, any k, and any $r \in [0,1]$. Note that because of M.2′, $(r, w, z), (r, w, z') \in \text{supp}(R_k, W, Z) = [0,1] \times \mathcal{WZ}$. Also because of M.2′ there exists a v_k^r such that $g_k(w, z, v_k^r) = Q_{X_k|WZ}(r|w, z)$ and $F_{X_k|WZ}\left(g_k(w, z, v_k^r) \mid w, z\right) = r$;

consider any such v_k^r.[35] Then

$$\begin{aligned} r = \mathbb{P}\left[X_k \leq g_k(w,z,v_k^r) \mid w,z\right] &= \mathbb{P}\left[g_k(w,z,V_k) \leq g_k(w,z,v_k^r) \mid w,z\right] \\ &= \mathbb{P}\left[g_k(w,z,V_k) \leq g_k(w,z,v_k^r) \,\middle|\, w,z'\right] \\ &= \mathbb{P}\left[g_k(w,z',V_k) \leq g_k(w,z',v_k^r) \,\middle|\, w,z'\right] \\ &= \mathbb{P}\left[X_k \leq g_k(w,z',v_k^r) \,\middle|\, w,z'\right], \end{aligned} \tag{22}$$

where the third equality is because $V_k \perp\!\!\!\perp Z|W$ by S.1 and the fourth equality is because $g_k(w,z,\cdot)$ and $g_k(w,z',\cdot)$ are strictly increasing by S.2. Assumption M.2′ and (22) show that $g_k(w,z,v_k^r) = Q_{X_k|WZ}(r \mid w,z)$ if and only if $g_k(w,z',v_k^r) = Q_{X_k|WZ}(r \mid w,z')$ for any v_k^r.

Now let $u \in [0,1]$ and consider

$$\begin{aligned} \mathbb{P}\left[U \leq u \mid w,z,R_k = r\right] &= \mathbb{P}\left[U \leq u \,\middle|\, w,z,X_k = Q_{X_k|WZ}(r \mid w,z)\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w,z,g_k(w,z,V_k) = Q_{X_k|WZ}(r \mid w,z)\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w,z',g_k(w,z,V_k) = Q_{X_k|WZ}(r \mid w,z)\right], \end{aligned}$$

since $U \perp\!\!\!\perp Z|W,V_k$ by S.1. As just shown, $[w,z',g_k(w,z,V_k) = Q_{X_k|WZ}(r \mid w,z)]$ and $[w,z',g_k(w,z',V_k) = Q_{X_k|WZ}(r \mid w,z')]$ are the same event, so

$$\begin{aligned} \mathbb{P}\left[U \leq u \mid w,z,R_k = r\right] &= \mathbb{P}\left[U \leq u \,\middle|\, w,z',g_k(w,z',V_k) = Q_{X_k|WZ}(r \mid w,z')\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w,z',X_k = Q_{X_k|WZ}(r \mid w,z')\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w,z',R_k = r\right]. \end{aligned}$$

This shows that $\mathbb{P}\left[U \leq u \mid w,z,R_k = r\right]$ does not depend on z, i.e. that $U \perp\!\!\!\perp Z|R_k,W$, and hence that CCI′ holds. Q.E.D.

Proof of Theorem 3. By Proposition 1′, it suffices to verify that $U \perp\!\!\!\perp Z|R_k,W$ for every k. Take any $u \in [0,1], r \in [0,1]$ and $(w,z),(w,z') \in \mathcal{WZ}$. Then

$$\begin{aligned} \mathbb{P}\left[U \leq u \mid w,z,R_k = r\right] &= \mathbb{P}\left[U \leq u \,\middle|\, w,z,F_{X_k|WZ}(X_k \mid w,z) = r\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w,z,F_{X_k(z)|W}(X_k(z) \mid w) = r\right], \end{aligned}$$

[35] S.2 implies that this v_k^r is unique, but this turns out not to be needed.

which follows by C.1 because for any x_k,

$$\begin{aligned} F_{X_k|WZ}(x_k \mid w, z) &= \mathbb{P}\left[X_k \leq x_k \mid w, z\right] \\ &= \mathbb{P}\left[X_k(z) \leq x_k \mid w, z\right] \\ &= \mathbb{P}\left[X_k(z) \leq x_k \mid w\right] = F_{X_k(z)|W}(x_k \mid w). \end{aligned}$$

Hence by C.2 and another application of C.1,

$$\begin{aligned} \mathbb{P}\left[U \leq u \mid w, z, R_k = r\right] &= \mathbb{P}\left[U \leq u \,\middle|\, w, z, F_{X_k(z')|W}(X_k(z') \mid w) = r\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w, z', F_{X_k(z')|W}(X_k(z') \mid w) = r\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w, z', F_{X_k|WZ}(X_k \mid w, z') = r\right] \\ &= \mathbb{P}\left[U \leq u \,\middle|\, w, z', R_k = r\right], \end{aligned}$$

which shows that $U \perp\!\!\!\perp Z | R_k, W$. Q.E.D.

Proof of Proposition 6. Assumption D.5 implies M.3. I will establish that $X = \rho(Z, U^d, U^s)$ for a function ρ such that (u^d, u^s) is strictly separable from z in ρ. Given D.5, this shows that S.1 and S.2′ are satisfied and hence that CCI holds by Theorem 2′.[36] I make use of the following property of twice differentiable functions with increasing first partial derivatives.

Proposition C.2. *Let $f : \mathcal{F} \subseteq \mathbb{R}^3 \to \mathbb{R}$ be a real-valued function with domain $\mathcal{F}$ that is twice continuously differentiable on* $\operatorname{int}\mathcal{F}$*. Suppose further that $\nabla_{x_1} f, \nabla_{x_2} f$ and $\nabla_{x_3} f$ are everywhere strictly positive on $\mathcal{F}$. Then (x_2, x_3) are strictly separable from x_1 in f if and only if for all $(x_1, x_2, x_3) \in \mathcal{F}$,*

$$\nabla_{x_1}\left(\frac{\nabla_{x_2} f(x_1, x_2, x_3)}{\nabla_{x_3} f(x_1, x_2, x_3)}\right) = 0.$$

Proof. See pp. 52-53 of Blackorby et al. (1978). The result is due to Leontief (1947) and Sono (1961). Q.E.D.

Under D.1-D.4 and M.1, Lemma 1 of Blundell and Matzkin (2010) shows that $P = \rho(Z, U^d, U^s)$ for some unique, twice continuously differentiable reduced form function

[36] Blundell and Matzkin (2010) show in their Appendix A.2 that strict separability of (u^d, u^s) from z in ρ implies their control function separability condition. They also show in their Theorem 1 that, given conditions slightly weaker than D.1-D.5, control function separability is sufficient and necessary for the existence of a first stage equation that satisfies Assumption S.

ρ that is strictly increasing in u^s.[37] For any $\bar{z}, \bar{u}^d, \bar{u}^s$, let $\bar{p} = \rho(\bar{z}, \bar{u}^d, \bar{u}^s)$ and $\bar{q} = d(\bar{p}, \bar{u}^d) = d(\rho(\bar{z}, \bar{u}^d, \bar{u}^s), \bar{u}^d)$. Then (13) implies

$$\rho(\bar{z}, \bar{u}^d, \bar{u}^s) = s\left(d\left[\rho(\bar{z}, \bar{u}^d, \bar{u}^s), \bar{u}^d\right], \bar{z}, \bar{u}^s\right). \tag{23}$$

The derivative of ρ with respect to z evaluated at $(\bar{z}, \bar{u}^d, \bar{u}^s)$ is found from (23) to be

$$\nabla_z \rho(\bar{z}, \bar{u}^d, \bar{u}^s) = \frac{\nabla_z s(\bar{q}, \bar{z}, \bar{u}^s)}{1 - \nabla_q s(\bar{q}, \bar{z}, \bar{u}^s)\nabla_p d(\bar{p}, \bar{u}^d)},$$

which is well-defined and strictly positive by D.4 and D.6. Similarly, differentiating (23) with respect to u^d yields

$$\nabla_{u^d} \rho(\bar{z}, \bar{u}^d, \bar{u}^s) = \frac{\nabla_q s(\bar{q}, \bar{z}, \bar{u}^s)\nabla_{u^d} d(\bar{p}, \bar{u}^d)}{1 - \nabla_q s(\bar{q}, \bar{z}, \bar{u}^s)\nabla_p d(\bar{p}, \bar{u}^d)}, \tag{24}$$

which is strictly positive by D.4 and M.1. It can also be seen that the derivative of ρ with respect to u^s is strictly positive (although this was already provided by Lemma 1 of Blundell and Matzkin (2010)),

$$\nabla_{u^s} \rho(\bar{z}, \bar{u}^d, \bar{u}^s) = \frac{\nabla_{u^s} s(\bar{q}, \bar{z}, \bar{u}^s)}{1 - \nabla_q s(\bar{q}, \bar{z}, \bar{u}^s)\nabla_p d(\bar{p}, \bar{u}^d)}, \tag{25}$$

since s is strictly increasing in u^s by D.3. Comparing the ratios of (24) and (25),

$$\begin{aligned}
\nabla_z \left(\frac{\nabla_{u^d} \rho(\bar{z}, \bar{u}^d, \bar{u}^s)}{\nabla_{u^s} \rho(\bar{z}, \bar{u}^d, \bar{u}^s)} \right) &= \nabla_z \left(\frac{\nabla_q s(\bar{q}, \bar{z}, \bar{u}^s)\nabla_{u^d} d(\bar{p}, \bar{u}^d)}{\nabla_{u^s} s(\bar{q}, \bar{z}, \bar{u}^s)} \right) \\
&= \nabla_{u^d} d(\bar{p}, \bar{u}^d) \nabla_z \left(\frac{\nabla_q s(\bar{q}, \bar{z}, \bar{u}^s)}{\nabla_{u^s} s(\bar{q}, \bar{z}, \bar{u}^s)} \right) = 0,
\end{aligned}$$

which follows from Proposition C.2, given D.3, D.4, D.6 and D.7. Since $\bar{z}, \bar{u}^d, \bar{u}^s$ were arbitrary, applying Proposition C.2 to ρ shows that (u^d, u^s) is strictly separable from z in ρ and hence that S.1 and S.2$'$ are satisfied. By Theorem 2$'$, this shows that CCI is satisfied.

Q.E.D.

[37] Blundell and Matzkin (2010) assume only once differentiability of s and d and the result of their lemma is that ρ is once differentiable. It is clear from their proof that higher order differentiability of s and d results in higher order differentiability of ρ.

D Proofs for Section 5

Notation. I let $\overline{\Gamma}_i(m)$ and $\hat{\Gamma}_i(m)$ stand for $\overline{\Gamma}(m, X_i, Z_i)$ (defined in the proof of Theorem 4) and $\hat{\Gamma}(m, X_i, Z_i)$. Given E.1, the i subscript is often unnecessary, so I also write $\overline{\Gamma}(m) = \overline{\Gamma}(m, x, z)$ and $\hat{\Gamma}(m) = \hat{\Gamma}(m, x, z)$ for arbitrary $(x, z) \in \overline{\mathcal{X}\mathcal{Z}}$. Finally, I often suppress the arguments of $\bar{q}(x|z) = Q_{\bar{X}|Z}(F_{\bar{X}}(x) \mid z)$ and write just $\bar{q}$ (similarly, $\hat{q}$), with the understanding that x, z are the same for all terms in the expression.

Throughout this section, B denotes an arbitrary, strictly positive constant which does not depend on the sample size or the parameter m.

Proof of Theorem 4. The estimator defined by (15) is an approximate sieve M-estimator. This is because

$$\begin{aligned}
\hat{T}(\hat{m}) &= \min_{m \in \mathcal{M}_n} \frac{1}{\left|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}\right|} \sum_{i \in \mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \left(\overline{\Gamma}_i(m) + \hat{\Gamma}_i(m) - \overline{\Gamma}_i(m) \right) \\
&\leq \min_{m \in \mathcal{M}_n} \frac{1}{\left|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}\right|} \sum_{i \in \mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \overline{\Gamma}_i(m) + \sup_{m \in \mathcal{M}_n} \frac{1}{\left|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}\right|} \sum_{i \in \mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \left| \hat{\Gamma}_i(m) - \overline{\Gamma}_i(m) \right| \\
&= \min_{m \in \mathcal{M}_n} \frac{1}{\left|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}\right|} \sum_{i \in \mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \overline{\Gamma}_i(m) + o_{\mathbb{P}}(1), \qquad (26)
\end{aligned}$$

where the uniform consistency of $\hat{\Gamma}$ for $\overline{\Gamma}$ is from Lemma D.1 below and $\overline{\Gamma}$ is defined as

$$\overline{\Gamma}(m, x, z) = \left\| Q_{Y|X} \left(F_{Y|XZ} \left(m \left[\bar{q}(x|z), \cdot \right] \mid \bar{q}(x|z), z \right) \mid x \right) - m(x, \cdot) \right\|_U^2 .$$

Note that the only difference between $\overline{\Gamma}$ and Γ (defined in (14)) is that $\overline{\Gamma}$ is evaluated at $\bar{q}$ instead of q. This is a result of the truncation discussed in the main text.

Sup-norm consistency of an approximate sieve M-estimator can be established by verifying Corollary 2.6 of White and Wooldridge (1991) or Theorem 3.1 of Chen (2007). I follow the latter and verify her Conditions 3.1-3.5.[38] In what follows, let $\overline{T}(m) = \mathbb{E}\,\overline{\Gamma}(m, \bar{X}, \bar{Z})$ denote the population average of $\overline{\Gamma}$. The average in (26) is the sample analog of $\overline{T}(m)$.

Condition 3.1 of Chen (2007) is the "identifiable uniqueness" condition of White and Wooldridge (1991) which says that the criterion function strictly identifies m_0, i.e. for any $\epsilon > 0$, $\inf_{m \in \mathcal{M}, \|m - m_0\|_\infty \geq \epsilon} \overline{T}(m) > \overline{T}(m_0)$. Since $\overline{T}$ is continuous in $\|\cdot\|_\infty$ (see below) and $(\mathcal{M}, \|\cdot\|_\infty)$ is compact by E.2, E.3 and the Arzéla-Ascoli Theorem, this condition is satisfied here with $\overline{T}(m_0) = 0$, due to Theorem 1. Condition 3.3 will also

[38] Since the setup of Chen (2007) is very general and allows for ill-posed estimation problems, many of these conditions will be redundant or trivially satisfied in my framework. This is primarily because her "$\delta(k(n))$" quantity can be taken as constant here.

be satisfied by the continuity of $\overline{T}$ since $k_n \to \infty$. Conditions 3.2 and 3.4 are general restrictions on the sieve spaces that are covered by E.6.

Condition 3.5 is uniform convergence in probability of the sample analog of $\overline{T}$ to $\overline{T}$. Since the data is i.i.d. by E.1 and $\overline{\Gamma}$ is bounded, pointwise convergence holds by a standard weak law of large numbers. Given the compactness of $(\mathcal{M}, \|\cdot\|_\infty)$, uniform convergence follows from Corollary 2.2 of Newey (1991) if $\left|\overline{\Gamma}(m) - \overline{\Gamma}(m')\right| \leq B \|m - m'\|_\infty$ for all $m, m' \in \mathcal{M}$ and a stochastically bounded random variable B. To see that this is satisfied, note that $\overline{\Gamma}^{1/2}$ is bounded (by 2), hence

$$\begin{aligned}
\left|\overline{\Gamma}(m) - \overline{\Gamma}(m')\right| &= \left|\left(\overline{\Gamma}(m)^{1/2} + \overline{\Gamma}(m')^{1/2}\right)\left(\overline{\Gamma}(m)^{1/2} - \overline{\Gamma}(m')^{1/2}\right)\right| \\
&\leq 4\Big| \left\|Q_{Y|X}\left(F_{Y|XZ}(m(\bar{q},\cdot) \mid \bar{q}, z) \mid x\right) - m(x,\cdot)\right\|_U \\
&\qquad - \left\|Q_{Y|X}\left(F_{Y|XZ}(m'(\bar{q},\cdot) \mid \bar{q}, z) \mid x\right) - m'(x,\cdot)\right\|_U \Big| \\
&\leq 4\Big\|Q_{Y|X}\left(F_{Y|XZ}(m(\bar{q},\cdot) \mid \bar{q}, z) \mid x\right) - m(x,\cdot) \\
&\qquad - Q_{Y|X}\left(F_{Y|XZ}(m'(\bar{q},\cdot) \mid \bar{q}, z) \mid x\right) + m'(x,\cdot)\Big\|_U \\
&\leq 4\Big\|Q_{Y|X}\left(F_{Y|XZ}(m(\bar{q},\cdot) \mid \bar{q}, z) \mid x\right) - \\
&\qquad - Q_{Y|X}\left(F_{Y|XZ}(m'(\bar{q},\cdot) \mid \bar{q}, z) \mid x\right)\Big\|_U + 4\left\|m(x,\cdot) - m'(x,\cdot)\right\|_U,
\end{aligned} \tag{27}$$

where the second inequality uses the reverse triangle inequality $|\|a\| - \|b\|| \leq \|a - b\|$ and the last inequality is the standard triangle inequality. Applying the mean value theorem to the first term twice,

$$(27)_{1^{\text{st}}} \leq \sup_{(y,x)\in\mathcal{Y}\overline{\mathcal{X}}} f_{Y|X}(y \mid x)^{-1} \sup_{(y,x,z)\in\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}} f_{Y|XZ}(y \mid x, z)\left\|m(\bar{q},\cdot) - m'(\bar{q},\cdot)\right\|_U.$$

From E.4, $f_{Y|X}$ is uniformly bounded away from 0 on $\mathcal{Y}\overline{\mathcal{X}}$ and $f_{Y|XZ}$ is uniformly bounded on $\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}$, so $(27)_{1^{\text{st}}} \leq B\|m(\bar{q},\cdot) - m'(\bar{q},\cdot)\|_U$. Since $\|m(x,\cdot) - m'(x,\cdot)\|_U$ and $\|m(\bar{q},\cdot) - m'(\bar{q},\cdot)\|_U$ are dominated by $\|m - m'\|_\infty$, conclude from (27) that $|\overline{\Gamma}(m) - \overline{\Gamma}(m')| \leq B\|m - m'\|_\infty$, as claimed. Note that this also shows that $\overline{T}$ is continuous in $\|\cdot\|_\infty$. Q.E.D.

Lemma D.1. *Under Assumptions E.1-E.5,*

$$\sup_{m\in\mathcal{M}_n} \frac{1}{\left|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}\right|} \sum_{i\in\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \left|\hat{\Gamma}_i(m) - \overline{\Gamma}_i(m)\right| = o_{\mathbb{P}}(1).$$

Proof of Lemma D.1. The first part of the argument is similar to that in Theorem 4; since $\hat{\Gamma}^{1/2}$ and $\overline{\Gamma}^{1/2}$ are bounded (by 2), $\left|\hat{\Gamma}(m) - \overline{\Gamma}(m)\right| \leq 4\left(\hat{\Gamma}(m)^{1/2} - \overline{\Gamma}(m)^{1/2}\right)$,

so from the reverse triangle inequality $|\|a\| - \|b\|| \leq \|a - b\|$,

$$\left|\hat{\Gamma}(m) - \overline{\Gamma}(m)\right| \leq 4\Big\|\hat{Q}_{Y|X}\left(\hat{F}_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right) \,\middle|\, x\right) - Q_{Y|X}\left(F_{Y|XZ}\left(m(\bar{q},\cdot) \mid \bar{q}, z\right) \mid x\right)\Big\|_U. \tag{28}$$

The rest of the proof shows that the upper bound in (28) is $o_{\mathbb{P}}(1)$ uniformly over $m \in \mathcal{M}_n$ and $(x,z) \in \overline{\mathcal{X}\mathcal{Z}}$.

Adding and subtracting $Q_{Y|X}\left(\hat{F}_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right) \,\middle|\, x\right)$ inside the norm in (28) and applying the triangle inequality leads to

$$\begin{aligned}\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}(28) \leq{}& 4\sup_{t\in(0,1)}\sup_{x\in\overline{\mathcal{X}}}\left|\hat{Q}_{Y|X}(t \mid x) - Q_{Y|X}(t \mid x)\right| \\ &+ 4\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\Big\|Q_{Y|X}\left(\hat{F}_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right) \,\middle|\, x\right) \\ &\qquad - Q_{Y|X}\left(F_{Y|XZ}\left(m(\bar{q},\cdot) \mid \bar{q}, z\right) \mid x\right)\Big\|_U.\end{aligned} \tag{29}$$

The first term of (29) is $o_{\mathbb{P}}(1)$ by Lemma D.2. Since $f_{Y|X}$ is uniformly bounded away from 0 by E.4, the derivative of $Q_{Y|X}$ is uniformly bounded above and hence the mean value theorem applied to the second term of (29) yields

$$(29)_{2^{\text{nd}}} \leq B\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left\|\hat{F}_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right) - F_{Y|XZ}\left(m(\bar{q},\cdot) \mid \bar{q}, z\right)\right\|_U. \tag{30}$$

Adding and subtracting $F_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right)$ inside the norm in (30) gives

$$\begin{aligned}(30) \leq{}& B\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left\|\hat{F}_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right) - F_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right)\right\|_U \\ &+ B\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left\|F_{Y|XZ}\left(m(\hat{q},\cdot) \mid \hat{q}, z\right) - F_{Y|XZ}\left(m(\bar{q},\cdot) \mid \bar{q}, z\right)\right\|_U.\end{aligned} \tag{31}$$

The first term of (31) is $o_{\mathbb{P}}(1)$ by Lemma D.2.

The second term of (31) can itself be broken up into two parts by adding and subtracting $F_{Y|XZ}\left(m(\hat{q},\cdot) \mid \bar{q}, z\right)$ so that

$$\begin{aligned}(31)_{2^{\text{nd}}} \leq{}& B\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left\|F_{Y|XZ}(m(\hat{q},\cdot) \mid \hat{q}, z) - F_{Y|XZ}(m(\hat{q},\cdot) \mid \bar{q}, z)\right\|_U \\ &+ B\sup_{m\in\mathcal{M}_n}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left\|F_{Y|XZ}(m(\hat{q},\cdot) \mid \bar{q}, z) - F_{Y|XZ}(m(\bar{q},\cdot) \mid \bar{q}, z)\right\|_U.\end{aligned} \tag{32}$$

I apply the mean value theorem to both terms in (32) in turn. For the first, note that $\nabla_x F_{Y|XZ}$ is uniformly bounded on $\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}$. This follows because $\nabla_x F_{Y|XZ}$ is contin-

uous by E.4 and $\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}$ is compact from (1), the compactness of $\overline{\mathcal{X}\mathcal{Z}}$ in E.2 and the smoothness of m in E.3. Then

$$\begin{aligned}(32)_{1^{\text{st}}} &\leq B \sup_{(y,x,z)\in\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}} \left|F_{Y|XZ}(y \mid \hat{q}, z) - F_{Y|XZ}(y \mid \bar{q}, z)\right| \\ &\leq B \sup_{(y,x,z)\in\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}} \left|\nabla_x F_{Y|XZ}(y \mid x, z)\right| \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} |\hat{q}(x|z) - \bar{q}(x|z)| = o_{\mathbb{P}}(1),\end{aligned}$$

from Lemma D.2. For the second term of (32),

$$\begin{aligned}(32)_{2^{\text{nd}}} &\leq B \sup_{(y,x,z)\in\mathcal{Y}\overline{\mathcal{X}\mathcal{Z}}} f_{Y|XZ}(y \mid x, z) \sup_{u\in[0,1]} \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} |m(\hat{q}, u) - m(\bar{q}, u)| \\ &\leq B \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} |\hat{q}(x|z) - \bar{q}(x|z)| = o_{\mathbb{P}}(1),\end{aligned}$$

where the bound on $f_{Y|XZ}$ is from E.4, the second inequality uses the Lipschitz continuity of m from E.3, and the uniform consistency of $\hat{q}$ for $\bar{q}$ is from Lemma D.2.

To summarize in reverse order, $(31)_{2^{\text{nd}}}$ is bounded by two $o_{\mathbb{P}}(1)$ terms and $(31)_{1^{\text{st}}}$ is $o_{\mathbb{P}}(1)$, hence (30) and (29) are $o_{\mathbb{P}}(1)$ as well. Returning to (28), conclude that

$$\sup_{m\in\mathcal{M}_n} \frac{1}{|\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}|} \sum_{i\in\mathcal{I}_{\overline{\mathcal{X}\mathcal{Z}}}} \left(\hat{\Gamma}_i(m) - \overline{\Gamma}_i(m)\right) = o_{\mathbb{P}}(1),$$

as claimed.

Q.E.D.

Lemma D.2. *Suppose Assumptions E.1-E.5 hold. Then*

$$\sup_{y\in\mathbb{R}} \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} \left|\hat{F}_{Y|XZ}(y \mid x, z) - F_{Y|XZ}(y \mid x, z)\right| = o_{\mathbb{P}}(1), \tag{33}$$

$$\sup_{t\in(0,1)} \sup_{x\in\overline{\mathcal{X}}} \left|\hat{Q}_{Y|X}(t \mid x) - Q_{Y|X}(t \mid x)\right| = o_{\mathbb{P}}(1), \tag{34}$$

$$\textit{and} \quad \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} |\hat{q}(x|z) - \bar{q}(x|z)| = o_{\mathbb{P}}(1). \tag{35}$$

Proof. By E.2 and E.4, there exists a compact set that strictly contains $\overline{\mathcal{X}}$ on which f_X is strictly positive. Similarly, if $\mathcal{Z}$ is not finite then there exists a compact set that strictly contains $\overline{\mathcal{Z}}$ on which f_Z is strictly positive and a compact set strictly containing $\overline{\mathcal{X}} \times \overline{\mathcal{Z}}$ on which f_{XZ} is strictly positive. (If $\mathcal{Z}$ is finite, no qualification is needed.)

Einmahl and Mason (2005) show in their Corollary 3 that given E.4 and E.5,[39]

$$\sup_{y\in\mathbb{R}}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left|\hat{F}_{Y|XZ}(y\mid x,z)-F_{Y|XZ}(y\mid x,z)\right|=o_{\mathbb{P}}(1), \tag{33}$$

$$\sup_{y\in\mathbb{R}}\sup_{x\in\overline{\mathcal{X}}}\left|\hat{F}_{Y|X}(y\mid x)-F_{Y|X}(y\mid x)\right|=o_{\mathbb{P}}(1), \tag{36}$$

$$\text{and}\quad \sup_{x\in\mathbb{R}}\sup_{z\in\overline{\mathcal{Z}}}\left|\hat{F}_{\bar{X}|Z}(x\mid z)-F_{\bar{X}|Z}(x\mid z)\right|=o_{\mathbb{P}}(1). \tag{37}$$

Since $\mathcal{Y}\overline{\mathcal{X}}$ is compact by (1), E.2 and E.3, and since $f_{Y|X}(\cdot\mid x)$ is bounded away from 0 for any $x\in\overline{\mathcal{X}}$ by E.4, the mapping $\hat{F}_{Y|X}(\cdot\mid x)\mapsto\hat{Q}_{Y|X}(\cdot\mid x)$ is continuous at $F_{Y|X}(\cdot\mid x)$ in the space of weakly increasing functions with the uniform metric.[40] Since $\hat{Q}_{Y|X}$ and $Q_{Y|X}$ are also continuous in x, it follows from an extension to the continuous mapping theorem (e.g., Theorem 7.25 of Kosorok (2008)) and (36) that

$$\sup_{t\in(0,1)}\sup_{x\in\overline{\mathcal{X}}}\left|\hat{Q}_{Y|X}(t\mid x)-Q_{Y|X}(t\mid x)\right|=o_{\mathbb{P}}(1),$$

which is (34). An analogous argument and (37) provides

$$\sup_{t\in(0,1)}\sup_{z\in\overline{\mathcal{Z}}}\left|\hat{Q}_{\bar{X}|Z}(t\mid z)-Q_{\bar{X}|Z}(t\mid z)\right|=o_{\mathbb{P}}(1), \tag{38}$$

noting that if $\mathcal{Z}$ is finite then the argument still holds, although $F_{\bar{X}|Z}$ is not continuous in z in that case.

It remains to show (35). First,

$$\begin{aligned}\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}|\hat{q}(x|z)-\bar{q}(x|z)|\leq&\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left|\hat{Q}_{\bar{X}|Z}(\hat{F}_{\bar{X}}(x)\mid z)-Q_{\bar{X}|Z}(\hat{F}_{\bar{X}}(x)\mid z)\right|\\&+\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}}\left|Q_{\bar{X}|Z}(\hat{F}_{\bar{X}}(x)\mid z)-Q_{\bar{X}|Z}(F_{\bar{X}}(x)\mid z)\right|.\end{aligned} \tag{39}$$

The first term is $o_{\mathbb{P}}(1)$ from (38). An application of the mean value theorem to the

[39] Einmahl and Mason (2005) also establish convergence rates for (33), (36) and (37). These can be useful for deriving a rate of convergence for the estimator in this paper. I note that the results of Einmahl and Mason (2005) are much stronger than used here and in particular are also uniform over general data-driven sequences of bandwidths.

[40] As is well-known (e.g., Section 3.9.4.2 of van der Vaart and Wellner (1996)), this mapping is actually Hadamard differentiable, but the differentiability is not needed here.

second term of (39) gives

$$\sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} \left| Q_{\bar{X}|Z}(\hat{F}_{\bar{X}}(x) \mid z) - Q_{\bar{X}|Z}(F_{\bar{X}}(x) \mid z) \right| \leq \sup_{(x,z)\in\overline{\mathcal{X}\mathcal{Z}}} f_{\bar{X}|Z}(x \mid z)^{-1} \sup_{x\in\mathbb{R}} \left| \hat{F}_{\bar{X}}(x) - F_{\bar{X}}(x) \right|,$$

which is $o_{\mathbb{P}}(1)$ since $f_{\bar{X}|Z}(x|z)$ is uniformly bounded away from 0 on $\overline{\mathcal{X}\mathcal{Z}} \subseteq \overline{\mathcal{X}}\times\overline{\mathcal{Z}}$ by E.4 and $\sup_{x\in\mathbb{R}} \left| \hat{F}_{\bar{X}}(x) - F_{\bar{X}}(x) \right| = o_{\mathbb{P}}(1)$ is the classic result of Glivenko-Cantelli. Q.E.D.

References

Abbring, J. H. and G. J. van den Berg (2003): "The Identifiability of the Mixed Proportional Hazards Competing Risks Model," *Journal of the Royal Statistical Society B*, 65, 701–710. 8

——— (2005): "Social Experiments and Instrumental Variables with Duration Outcomes," *Tinbergen Institute Discussion Paper TI-2005-047/3*. 8

Altonji, J. G. and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogeneous Regressors," *Econometrica*, 73, 1053–1102. 3

Angrist, J., K. Graddy, and 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. 26

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442. 21

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. 21

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014. 14

Athey, S. and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-In-Differences Models," *Econometrica*, 74, 431–497. 3, 8

Benkard, C. L. and S. Berry (2006): "On the Nonparametric Identification of Nonlinear Simultaneous Equations Models: Comment on Brown (1983) and Roehrig (1988)," *Econometrica*, 74, 1429–1440. 25

Blackorby, C., D. Primont, and R. R. Russell (1978): *Duality, Separability, and Functional Structure*, Elsevier North-Holland, Inc. 19, 47

Blundell, R. and R. L. Matzkin (2010): "Conditions for the Existence of Control Functions in Nonseparable Simultaneous Equations Models," *cemmap working paper CWP28/10*. 25, 26, 47, 48

Blundell, R. and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky, Cambridge: Cambridge University Press, vol. II, 312–357. 9

Bond, S. J. and J. E. H. Shaw (2006): "Bounds on the Covariate-Time Transformation for Competing-Risks Survival Analysis," *Lifetime Data Analysis*, 12, 285–303. 8

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. N. Christofides, K. E. Grant, and R. Swidinsky, Toronto: University of Toronto Press, 201–222. 24

Chaudhuri, P. (1991): "Nonparametric Estimates of Regression Quantiles and their Local Bahadur Representation," *The Annals of Statistics*, 19, 760–777. 7

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," *Handbook of Econometrics*, 6B. 27, 29, 30, 49

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. 3, 4, 6, 7, 8, 21

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4–14. 3, 4, 7

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. 7

——— (2005): "Nonparametric Identification Under Discrete Variation," *Econometrica*, 73, 1525–1550. 14

——— (2007): "Instrumental Values," *Journal of Econometrics*, 139, 15–34. 7, 8, 14

Doksum, K. (1974): "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-sample Case," *The Annals of Statistics*, 2, 267–277. 6, 7, 20

Duflo, E. (2001): "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *The American Economic Review*, 91, 795–813. 4, 23

Einmahl, U. and D. M. Mason (2000): "An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators," *Journal of Theoretical Probability*, 13. 29

——— (2005): "Uniform in Bandwidth Consistency of Kernel-Type Function Estimators," *The Annals of Statistics*, 33, 1380–1403. 29, 53

Evdokimov, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," *working paper*. 3

Feinstein, L. and J. Symons (1999): "Attainment in Secondary School," *Oxford Economic Papers*, 51, 300–321. 4, 21

Florens, J., J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogeneous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 3

Härdle, W., P. Janssen, and R. Serfling (1988): "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *The Annals of Statistics*, 16, 1428–1449. 29

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. 3

——— (1990): "Selectivity Bias: New Developments," *The American Economic Review*, 80, 313–318. 3

——— (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748. 2

Heckman, J. J. and B. E. Honoré (1989): "The Identifiability of the Competing Risks Model," *Biometrika*, 76, 325–330. 8

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535. 2, 8, 21

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *The Review of Economics and Statistics*, 88, 389–432. 2

Heckman, J. J. and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, 669–738. 2

Hoderlein, S. and E. Mammen (2009): "Identification and Estimation of Local Average Derivatives in Non-Separable Models Without Monotonicity," *Econometrics Journal*, 12, 1–25. 7

Hoxby, C. M. (2000): "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *The Quarterly Journal of Economics*, 115, 1239–1285. 4, 22

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 3, 20, 21, 26

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 3, 7, 14, 19, 24

Kasy, M. (2011): "Identification in Triangular Systems Using Control Functions," *forthcoming in Econometric Theory*, 27. 19

Koenker, R. (2005): *Quantile Regression*, Cambridge University Press. 2

Komunjer, I. and A. Santos (2010): "Semiparametric Estimation of Nonseparable Models: A Minimum Distance From Independence Approach," *Working paper*. 13

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*, Springer: New York. 53

Lehmann, E. (1974): *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, Inc. 6

Leontief, W. (1947): "A Note on the Interrelation of Subsets of Independent Variables of a Continuous Function with Continuous First Derivatives," *Bulletin of the American Mathematical Society*, 53, 343–350. 19, 47

Manski, C. F. (1983): "Closest Empirical Distribution Estimation," *Econometrica*, 51, 305–319. 13

Matzkin, R. L. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375. 4, 6, 7, 11, 34, 35

——— (2008): "Identification in Nonparametric Simultaneous Equations Models," *Econometrica*, 76, 945–978. 25

Nelson, R. B. (2006): *An Introduction to Copulas*, Springer, 2nd ed. 5

Newey, W. K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161–1167. 50

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578. 7

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603. 19

Patton, A. J. (2006): "Modelling Asymmetric Exchange Rate Dependence," *International Economic Review*, 47, 527–556. 5

Rényi, A. (1970): *Foundations of Probability*, Holden-Day. 9

Sklar, A. (1959): "Fonctions de rèpartition á n dimensions et leurs marges," *Publications de l'Institut Statistique de l'Universitè de Paris*, 8, 229–231. 5

Sono, M. (1961): "The Effect of Price Changes on the Demand and Supply of Separable Goods," *International Economic Review*, 2, 239–271. 47

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer-Verlag. 53

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341. 21

White, H. and J. M. Wooldridge (1991): "Results on Sieve Estimation with Dependent Observations," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen. 29, 49

			$N = 200$			$N = 400$			$N = 800$		
μ_Z	F_Z	corr(X, Z)	bias	(std)	MSE	bias	(std)	MSE	bias	(std)	MSE
	$\mathrm{B}\left(\frac{1}{2}\right)$	.453	.016	(.171)	.030	.009	(.116)	.014	.017	(.086)	.008
1	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.420	-.129	(.195)	.055	-.084	(.141)	.027	-.075	(.105)	.017
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.254	-.210	(.344)	.163	-.135	(.235)	.074	-.119	(.174)	.045
	$\mathrm{B}\left(\frac{1}{2}\right)$	.256	.016	(.291)	.085	-.017	(.211)	.045	-.008	(.156)	.024
.5	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.218	-.205	(.397)	.200	-.141	(.258)	.086	-.125	(.191)	.052
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.223	-.126	(.571)	.342	-.138	(.461)	.231	-.134	(.357)	.146
	$\mathrm{B}\left(\frac{1}{2}\right)$	.143	.191	(.430)	.222	.077	(.345)	.125	.040	(.262)	.070
.25	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.102	-.003	(.570)	.325	-.056	(.473)	.227	-.128	(.392)	.170
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.002	.150	(.709)	.525	.171	(.675)	.485	-.022	(.568)	.323

Table 1: (MC1): $Y = \theta(X - 10) + XU/10; \quad \theta_0 = 1.$

			$N = 200$			$N = 400$			$N = 800$		
μ_Z	F_Z	corr(X, Z)	bias	(std)	MSE	bias	(std)	MSE	bias	(std)	MSE
	$\mathrm{B}\left(\frac{1}{2}\right)$	.453	-.072	(.066)	.010	-.059	(.049)	.006	-.050	(.037)	.004
1	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.420	-.118	(.065)	.018	-.095	(.047)	.011	-.081	(.034)	.008
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.254	-.140	(.105)	.031	-.108	(.071)	.017	-.086	(.053)	.010
	$\mathrm{B}\left(\frac{1}{2}\right)$	.256	-.110	(.106)	.023	-.084	(.083)	.014	-.063	(.060)	.008
.5	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.218	-.148	(.121)	.036	-.116	(.078)	.020	-.091	(.058)	.012
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.223	-.199	(.148)	.062	-.149	(.123)	.037	-.113	(.091)	.021
	$\mathrm{B}\left(\frac{1}{2}\right)$	.143	-.160	(.133)	.043	-.130	(.113)	.030	-.095	(.089)	.017
.25	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.102	-.202	(.141)	.061	-.159	(.133)	.043	-.119	(.101)	.024
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.002	-.237	(.150)	.079	-.199	(.148)	.062	-.155	(.155)	.048

Table 2: (MC2): $Y = \theta(X - 10)^2 + XU/10; \quad \theta_0 = .3.$

			$N = 200$			$N = 400$			$N = 800$		
μ_Z	F_Z	corr(X, Z)	bias	(std)	MSE	bias	(std)	MSE	bias	(std)	MSE
	$\mathrm{B}\left(\frac{1}{2}\right)$	.453	-.054	(.065)	.007	-.044	(.050)	.004	-.037	(.040)	.003
1	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.420	-.126	(.079)	.022	-.100	(.054)	.013	-.087	(.046)	.010
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.254	-.161	(.123)	.041	-.116	(.082)	.020	-.096	(.061)	.013
	$\mathrm{B}\left(\frac{1}{2}\right)$	.256	-.104	(.124)	.026	-.083	(.089)	.015	-.062	(.068)	.008
.5	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.218	-.191	(.147)	.058	-.135	(.095)	.027	-.111	(.069)	.017
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.223	-.249	(.205)	.104	-.178	(.161)	.058	-.139	(.112)	.032
	$\mathrm{B}\left(\frac{1}{2}\right)$	.143	-.158	(.192)	.062	-.130	(.143)	.037	-.098	(.107)	.021
.25	$\mathrm{U}\left(\frac{1\pm\sqrt{3}}{2}\right)$	.102	-.243	(.196)	.097	-.187	(.186)	.069	-.149	(.121)	.037
	$N\left(\frac{1}{2}, \frac{1}{4}\right)$	.002	-.291	(.235)	.140	-.221	(.221)	.098	-.196	(.201)	.079

Table 3: (MC3): $Y = \theta(X - 10)^2 + \theta^2(X - 10) + XU/10; \quad \theta_0 = .5.$