Review of Economic Studies (2014) **81**, 1614–1636 doi:10.1093/restud/rdu018

Advance access publication 28 May 2014

Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment

MAXIMILIAN KASY

Department of Economics, Harvard University, and IHS Vienna

First version received January 2012*; final version accepted April* 2014 (*Eds.*)

This article discusses identification in continuous triangular systems without restrictions on heterogeneity or functional form. We do not assume separability of structural functions, restrictions on the dimensionality of unobservables, or monotonicity in unobservables. We do maintain monotonicity of the first stage relationship in the instrument and consider the case of real-valued treatment. Under these conditions alone, and given rich enough support of the data, potential outcome distributions, the average structural function, and quantile structural functions are point identified. If the support of the continuous instrument is not large enough, potential outcome distributions are partially identified. If the instrument is discrete, identification fails completely. If treatment is multi-dimensional, additional exclusion restrictions yield identification.

The set-up discussed in this article covers important cases not covered by existing approaches such as conditional moment restrictions (cf. Newey and Powell, 2003) and control variables (cf. Imbens and Newey, 2009). It covers, in particular, random coefficient models, as well as systems of structural equations.

Key words: Triangular systems, Simultaneous equations, Instrumental variables, Identification

JEL Codes: C14, C26, C31, C36

1. INTRODUCTION

A large literature in econometrics studies identification of triangular systems of the form

$$Y = g(X, U)$$
$$X = h(Z, V),$$

where it is assumed that $Z \perp (U, V)$. The variable Y is usually called the outcome variable, X is the treatment, and Z is an instrumental variable. U and V are unobservable. Such a triangular system structure is assumed in most discussions of instrumental variable methods and their non-parametric generalizations. Triangular systems generalize the notion of a randomized experiment, which corresponds to the case where V is constant, and are thus closely related to our notion of causality. The basic idea of all approaches to identification in triangular systems is that variation in X which is induced by variation in Z is random relative to the unobserved U, and can thus be used to learn something about the distribution of counterfactual outcomes for Y and about causal effects.

In this article we discuss triangular systems where both X and Z are real-valued and continuously distributed, while the heterogeneity terms U and V are allowed to be of arbitrary dimension. They may enter the functions g and h in unrestricted ways. This case of unrestricted heterogeneity has not been considered in the literature before. We do assume monotonicity of the function h in the instrument Z. Under this condition, and with sufficient support of the data, we can show point identification of the distribution of potential outcomes $Y^x = g(x, V)$, and thus in particular of the average structural function (ASF) and the quantile structural functions (QSF). The intuition of our main identification result can be described as follows: if we assume monotonicity and sufficient support, then there is exactly one value Z^x for each member of the population such that this member would receive treatment $X = x$ if $Z = Z^x$. As we move Z around, we thus observe treatment $X = x$ and the corresponding potential outcome Y^x exactly once for everyone. This suggests that we might only need to reweight across values of Z given X to get the distribution of Y^x for the entire population. Our proof of identification shows that this intuition is correct.

Our set-up generalizes the models considered in the literature on conditional moment restrictions (cf. Newey and Powell, 2003) which restrict heterogeneity in the structural equation of interest. Our set-up also generalizes the models considered in the literature on control variables (cf. Imbens and Newey, 2009) which restrict heterogeneity in the first-stage relationship. To the best of our knowledge, this article is the first to show point identification in continuous triangular systems without restrictions on heterogeneity.

There are a number of reasons why the literature has focused on generalizing models of triangular systems to allow for multi-dimensional heterogeneity, either in the first stage or the structural equation of interest, and which motivate interest in the case of unrestricted heterogeneity in this article: (i) One-dimensional heterogeneity implies that there is no heterogeneity in causal effects given outcome levels, an implication that is empirically wrong in many contexts, as shown for instance by Heckman *et al.* (1997). (ii) Economic models suggest there might be heterogeneity in preferences, technologies, endowments, beliefs, measurement error, etc., and each of these is likely to be multi-dimensional in turn. In Section 3.1, we discuss several examples of economically motivated models that lead to multi-dimensional heterogeneity in both the first stage and the structural equation of interest. (iii) One-dimensional heterogeneity in U (and invertibility of g in U) implies that we can predict *all* counterfactual outcomes for Y as we change X, based on observation of one realization of (X, Y) alone; for any x, Y^x is a function of (X, Y). The same holds for first-stage heterogeneity V, treatment X, and instrument Z. (iv) The assumption that U is unidimensional has strong identifying power. This assumption is not without loss of generality in settings where the *joint distribution* of Y^x across different values of x matters, which is the case in triangular systems when we use the non-parametric IV (conditional moment restriction) approach. Similarly, the assumption that V is unidimensional is not without loss of generality in settings where the *joint distribution* of X^z across different values of z matters, which is the case in triangular systems when we use the control function approach.

The rest of this article is structured as follows. Section 2 provides a brief literature review. Section 3 introduces the formal set-up analysed in this article, and discusses the assumptions maintained. Section 3.1 discusses some examples of economic models with multi-dimensional heterogeneity in both the first stage and the structural equation of interest, which are not covered by any of the existing approaches but do satisfy our assumptions.

Section 4 discusses the main results of this article. Section 4.1 presents our central positive result, characterizing identification under the assumption of monotonicity of h in Z. Section 4.2 provides sharp bounds for the case that the support of Z is insufficient for point identification. Section 4.3 discusses the relationship of our identification results to the control function approach. Section 4.4 shows non-identification for the case of discrete Z. Section 5 sketches a reweighting estimator based on the identification result of Section 4.1, and briefly discusses the properties

of such estimators and issues of inference. (We do not fully characterize asymptotic properties or rates of convergence, however.) Section 6 discusses extensions to higher dimensions. This section shows in particular how point identification can be achieved when both X and Z are k-dimensional, if additional exclusion restrictions are imposed on the first-stage relationship. Section 7 concludes. Appendix A contains all proofs.

A Supplementary Appendix provides additional results and discussions. Supplementary Appendix A shows that the triangular system we consider arises naturally as reduced form of a system of simultaneous equations. Supplementary Appendix B interprets partial- and non-identification in our context as arising from support problems. Supplementary Appendix C discusses the construction of estimators of minimal variance in overidentified settings. Supplementary Appendix D provides a numerical illustration of our main results, motivated by the returns to education model discussed in Card (2001). This appendix also demonstrates the finite sample performance of reweighting estimators and control function estimators in a series of Monte Carlo simulations.

2. A BRIEF LITERATURE REVIEW

Identification of triangular systems with unrestricted heterogeneity for binary (discrete) treatment X and instrument Z is well understood, in particular since the contributions of Imbens and Angrist (1994), Manski (2003), and Heckman and Vytlacil (2005). Identification of triangular systems with continuous treatment X is the subject of more recent contributions to the literature.

There are two main approaches in the literature on triangular systems with continuous treatment. One strand of the literature assumes an additively separable structural relationship of the form $Y=g^1(X)+U$, where U denotes one-dimensional unobserved heterogeneity. Under this assumption, the ASF g^1 satisfies the conditional moment restriction $E[Y-g^1(X)|Z]=0$ (cf. Newey and Powell, 2003; Horowitz, 2011). Hahn and Ridder (2011) show that, absent additivity, the function identified by the conditional moment restriction has no structural interpretation. For a general triangular system, the function g^1 identified by the conditional moment restriction satisfies

$$E[g^1(X)|Z]=E[Y|Z]=\int g(h(z,v),u)dP(u,v),$$

whereas the ASF $\overline{g}$ satisfies

$$E[\overline{g}(X)|Z]=\int g(h(z,v),u)dP(u)dP(v).$$

Chernozhukov and Hansen (2005) and Chernozhukov *et al.* (2007) generalize the conditional moment restriction approach to structural equations monotonic in unobservables. Their approach also relies on the assumption of one-dimensional heterogeneity U, while dropping the requirement of additive separability. Chernozhukov and Hansen (2005) furthermore discuss a slightly more general "rank similarity" condition, which allows for random (exogenous) noise in the outcome equation, in addition to one-dimensional unobserved heterogeneity. Jun *et al.* (2011) provide set identification results in a set-up similar to Chernozhukov and Hansen (2005) and Chernozhukov *et al.* (2007), and building on Chesher (2003, 2005), with one-dimensional heterogeneity in the structural equation of interest, imposing some additional conditions.

Another strand of the literature uses control functions (cf. Imbens and Newey, 2009; Newey *et al.*, 1999). Imbens and Newey (2009) propose the control function $V^*=F_{X|Z}(X|Z)$. They show that conditional independence $X \perp U|V^*$ holds if h is monotonic in V, where V is

one-dimensional. Kasy (2011) shows that conditional independence fails in general if V is of higher dimension: if V is more than one-dimensional, then the family of conditional distributions $P(U|X,Z)$ will in general be two-dimensional. As a consequence there is no control function $C(X,Z)$ such that (i) $P(U|X=x,C(X,Z)=c)$ does not depend on x given c, and (ii) X has non-degenerate support given $C(X,Z)$. It turns out, however, that the approach proposed by Imbens and Newey (2009) does identify the ASF under the assumptions maintained in this article. This is quite surprising given the failure of conditional independence. Florens *et al.* (2008) propose a control function approach similar to Imbens and Newey (2009). In addition to the assumptions of Imbens and Newey (2009), they impose the assumption of a stochastic polynomial functional form for the structural equation of interest, which allows to extrapolate outside the support of the data and to achieve point identification even when full support is absent.

There is, finally, the literature on non-parametric identification in systems of simultaneous equations, see in particular the important contributions by Chesher (2003) and Matzkin (2008). This literature requires that the dimensionality of unobserved heterogeneity is no larger than the number of endogenous variables. The simultaneous systems considered in the present article, in contrast, allow for unrestricted heterogeneity. Blundell and Matzkin (2010) provide conditions under which simultaneous systems have a triangular reduced form with one-dimensional heterogeneity in the first stage.

Full heterogeneity of structural functions, as in the model considered here, is allowed in some contributions to the literature assuming exogeneity of X, for instance in Hoderlein and Mammen (2007). The importance of multi-dimensional heterogeneity has been emphasized in many contributions to the literature. Heckman *et al.* (1997) "present evidence that heterogeneity in response to programmes is empirically important". Chernozhukov and Hansen (2005) note that "rank invariance implies that the potential outcomes Y^d are not truly multivariate, being jointly degenerate, which may be implausible on logical grounds". Imbens (2007), discussing the control function approach, states that "such a triangular system corresponds to a special and potentially restrictive form of endogeneity, especially with a single unobserved component combined with monotonicity in the choice equation", and that "the assumption of a scalar unobserved component in the choice equation that enters in a monotone way is very informative".

Many more recent contributions would deserve a review, let me just mention a few: Hoderlein and Sasaki (2013) consider a framework similar to Imbens and Newey (2009), where first stage heterogeneity is assumed to be one-dimensional and the first-stage is monotonic in the heterogeneity term. In contrast to the present article, which is interested in the distribution of potential outcomes, they consider "outcome conditional policy effects", building on the insights of Hoderlein and Mammen (2007). Schennach *et al.* (2012) discuss a setting very similar to ours, where heterogeneity of arbitrary dimension is allowed to enter the structural relationships in a non-separable way. They consider derivative ratio (local indirect least squares) estimators and conclude that in general such methods cannot recover the average marginal effect. They can, however, test the hypothesis of no effect. Han (2012) considers a model with the same identifying assumptions as Newey *et al.* (1999) (additively separable one-dimensional first-stage heterogeneity), and discusses inference under a form of weak instrument asymptotics, shrinking the first stage relationship to a constant function in the limit. In the recent literature on triangular systems with discrete treatments, Gautier and Hoderlein (2011), building on the insights of Hoderlein *et al.* (2010), discuss identification under the functional form restriction of a selection equation containing multiple unobservables that enter through a non-parametric random coefficients specification. Jun *et al.* (2010) discuss identification in a three equation triangular system with two endogenous discrete treatments and one-dimensional latent heterogeneity in each equation.

3. SET-UP

We consider the following set-up.

Assumption 1. (Triangular system)

$$Y = g(X, U)$$
$$X = h(Z, V), \tag{1}$$

where X, Y, Z are random variables taking their values in $\mathbb{R}$, the unobservables U, V have their support in an arbitrary measurable space of unrestricted dimensionality, and

$$Z \perp (U, V). \tag{2}$$

Assumption 1 states that the observed data (Y, X, Z) are generated by a triangular system which satisfies the usual properties that Z is (as if) randomly allocated, and that the only causal effect of Z on Y is through X (exclusion restriction). Assumption 1 additionally states that we are discussing the case of real-valued X and Z; extensions to higher dimensions will be discussed in Section 6.

Assumption 2. (Continuous treatment) *Treatment X is continuously distributed in $\mathbb{R}$ conditional on Z.*

Assumption 2 states that we are discussing the case of continuously distributed X, as opposed to the case of discrete treatments, which is different in many respects.

Assumption 3. (First-stage monotonic in instrument) *The first-stage relationship $h(z, v)$ is strictly increasing in z for all v.*

Assumption 3 states that the effect of the instrument Z on the treatment X goes in the same direction for every member of the population. This assumption is central to our argument and will be discussed in detail further.

Assumption 4. (Continuous instrument) *The instrument Z is continuously distributed in $\mathbb{R}$, with support $[z_l, z_u]$. The first-stage relationship h is continuous in z for all z and almost all v, and $P(X \leq x | Z = z)$ is continuous in z for all x.*

Assumption 4 states that the instrument is continuously distributed. The case of a discrete instrument is discussed in Section 4.4, where it is shown that in that case identification fails completely. Assumption 4 furthermore introduces the notation $[z_l, z_u]$ for the support of Z.

Note that we have left the support of Y unrestricted. Our set-up allows for binary outcomes Y, in particular.

Under these assumptions, we can introduce the following potential outcome notation.[1]

1. In this article, "structural functions" and "potential outcomes" are considered to be essentially equivalent notations for random functions, where the arguments are either written as function arguments or as superscripts. The structural function notation additionally makes explicit the unobserved heterogeneity, which is left implicit in the potential outcome notation.

Definition 1. (Potential outcomes) *We denote*

$$Y^x = g(x, U)$$
$$X^z = h(z, V). \tag{3}$$

We define furthermore

$$Z^x = \begin{cases} h^{-1}(x, V) & \text{if } h(z_l, V) \leq x \leq h(z_u, V) \\ -\infty & \text{if } x < h(z_l, V) \\ \infty & \text{if } h(z_u, V) < x. \end{cases} \tag{4}$$

The notation Y^x and X^z for potential outcomes is standard. The variable Z^x is a slightly different object. It denotes the value of Z which would attain treatment level x given unobserved heterogeneity V. In equation (4), Z^x is well defined under Assumption 3; h^{-1} is to be understood as the inverse of h given V. We set Z^x to equal $\pm\infty$ if there is no $z \in [z_l, z_u]$ that achieves $h(z, V) = x$.

The main object of interest in this article is the ASF $\overline{g}$ (cf. Blundell and Powell, 2003), where

$$\overline{g}(x) := E[Y^x] = E[g(x, U)]. \tag{5}$$

Our results discuss, more generally, identification of the (marginal) distribution of the potential outcome Y^x.

Remark: We have not imposed any of the restrictions on heterogeneity common in the literature. In particular, we have not assumed that V can be replaced by a one-dimensional index, as necessary for achieving conditional independence using the control function approach (cf. Imbens and Newey, 2009; Kasy, 2011). We have not assumed, either, that the structural function of interest is separable in U, as necessary for the "non-parametric instrumental variable" approach (cf. Hahn and Ridder, 2011; Horowitz, 2011; Newey and Powell, 2003). The assumptions maintained in this article thus cover, in particular, the case of a generic random coefficient model, which is not covered by either the control function or the conditional moment restriction literature. Sections 3.1 and Supplementary Appendix D discuss examples of such random coefficient models.

Remark: We are imposing one substantive restriction, monotonicity of h in Z. This assumption is potentially testable, since it implies monotonicity of $F_{X|Z}(X|Z)$ in Z for all X when Z is independent of V.[2] I would argue, also, that this assumption is more easily justified than monotonicity of h in V, as required for the control function approach if conditional independence is to be achieved. The latter assumption requires heterogeneity to be effectively one-dimensional. If V is one-dimensional, then we can predict the counterfactual X^z for any given unit of observation by knowing her realized X and Z. There is *no* unobserved heterogeneity given the observations.

A typical economic model that implies monotonicity in Z is as follows. Assume X is the solution to a utility maximization problem, and Z is a cost shifter,

$$X^z = \operatorname*{argmax}_x u(x, z, V), \tag{6}$$

2. This is necessary but not sufficient for our monotonicity condition to hold. Monotonicity of $F_{X|Z}(X|Z)$ is a condition on the marginal distributions of X^z, whereas our monotonicity requirement restricts the joint distribution of X^z across different values of z.

where u is utility (net of costs), and V captures heterogeneity in utility and costs. The first-order condition for this problem is $\partial u/\partial x = 0$, which implies

$$\frac{\partial X^z}{\partial z} = -\frac{\partial^2 u}{\partial x \partial z} \Big/ \frac{\partial^2 u}{\partial x^2}. \tag{7}$$

Monotonicity of h in z in this setting is guaranteed if (i) $\partial^2 u/\partial x \partial z > 0$ (an increase in z decreases the marginal cost of x), and (ii) $\partial^2 u \Big/ \partial x^2 < 0$ (decreasing marginal utility of x).

We can also compare monotonicity in Z and monotonicity in V in mathematical terms. The general triangular system model of Assumptions 1, 2, and 4 induces a probability distribution on the space of structural functions mapping Z to X. A realization $h(.,V)$ from this distribution describes the schedule of potential treatment levels X^z for a given unit of observation. Without further restrictions, this is a distribution on the infinite-dimensional space $\mathscr{C}([z_l, z_u])$, the space of continuous functions on the interval $[z_l, z_u]$. The assumption that h is monotonic in V implies that the support of this distribution is restricted to a one-dimensional sub-manifold of $\mathscr{C}([z_l, z_u])$. In contrast, the assumption that h is monotonic in z restricts the support of this distribution to a proper subset of $\mathscr{C}([z_l, z_u])$, but it does not restrict it to lie in a subset of lower dimension.

The assumption of monotonicity of h in Z, as maintained here, is similar to the "no defiers" assumption in the discrete Imbens and Angrist (1994) set-up.[3] Note that our monotonicity assumption is unrelated to the "monotone instrumental variables" assumption of Manski and Pepper (2003), who in fact *weaken* the exogeneity and exclusion restriction of the standard triangular system, whereas we are imposing an additional condition.

3.1. *Examples*

The analysis in the present article is motivated by the argument that, in general, it is unlikely *a priori* that heterogeneity in either the first stage or the structural equation of interest is one-dimensional. Economic models suggest that there might be heterogeneity in preferences, technologies, initial endowments, and beliefs, as well as measurement error, etc. Any of these components of heterogeneity might furthermore well be multi-dimensional, in turn. In this section, we discuss several stylized economic models which imply multi-dimensional heterogeneity in both the first stage and the structural equation, and which are thus not covered by any of the existing approaches in the literature, but are covered by the assumptions maintained in this article. The purpose of these examples is not so much to be taken literally as semi-parametric models of an actual data generating process, but rather to illustrate that multi-dimensional heterogeneity arises quite naturally in many economic settings.

3.1.1. Production function. The first example we consider is production function estimation, using random variation in either output or input prices.[4] Let Y denote log output of a firm, X log labour demand of this firm, Z log output price, and W log wage. We assume that firms produce according to a Cobb–Douglas production function, where output is given by $\exp(U_1 + U_2 X)$. Here U_1 depends on the amount of capital used by the firm and total factor productivity, and U_2, the elasticity of output with respect to labour, is smaller than 1 almost surely.

3. Some recent contributions study the implications of dropping the "no defiers" assumption in the Imbens and Angrist (1994) set-up, see in particular de Chaisemartin (2012) and Huber and Mellace (2010).

4. I thank Bryan Graham for suggesting this example to me.

This elasticity depends on the technology used by the firm. We get the firm-specific production function (in logarithms),

$$Y^x = g(x, U) = U_1 + U_2 \cdot x. \tag{8}$$

This is our structural equation of interest. Profits for the firm are given by

$$\pi(x) = e^Z \cdot e^{Y^x} - e^W \cdot e^x. \tag{9}$$

Assume that labour inputs are chosen by the firm to maximize profits, so that

$$X = \operatorname*{argmax}_x \pi(x). \tag{10}$$

The first-order condition for the optimal choice of X given $Z = z$, expressed in logarithms, yields

$$\log(U_2) + z + U_1 + U_2 \cdot X^z = W + X^z \tag{11}$$

and thus

$$X^z = h(z, V) = \frac{\log(U_2) + U_1}{1 - U_2} + \frac{1}{1 - U_2}(z - W). \tag{12}$$

If it can be argued that variation of the output price (or alternatively of wages) is random across firms, we can use the output price (or alternatively wages) as an instrument for labour when estimating the average log production function. Equation (12) is then the first-stage relationship of a triangular system which satisfies the assumptions maintained in this article. In this example, we assumed that there is heterogeneity in both capital inputs and technology employed by firms. Heterogeneity of these two components implies multi-dimensional heterogeneity in both the first stage and the equation of interest.

3.1.2. Returns to education. Next, consider the model of optimal schooling choice by individuals discussed by Card (2001). Let Y denote log earnings, X years of education, and Z distance to college. Card (2001) assumes that potential earnings given years of education $X = x$ take the form

$$Y^x = U_1 + U_2 \cdot x - \frac{k}{2} \cdot x^2 \tag{13}$$

for some constant k and heterogeneity in both the level of earnings U_1 and the returns to schooling U_2. Assume further that the marginal cost of schooling conditional on $Z = z$, which subsumes all considerations other than the economic returns to education, is given by

$$C(z) = V_1 + V_2 \cdot z + V_3 \cdot z^2, \tag{14}$$

where $V_2 > 0$ and $V_3 > 0$ almost surely. Heterogeneity in V_1 might for instance be driven by family background, heterogeneity in V_2 and V_3 by variations in the local and regional transport infrastructure connecting a potential student's house to the nearest college. Alternatively, Z might be a randomly allocated fellowship for going to college, and heterogeneity in V might be driven by a student's family's financial endowments, taste for education, and time preference. Solving for the optimal X given $Z = z$ which equates marginal returns to education, $U_2 - k \cdot x$, to marginal costs, we get

$$X^z = \frac{1}{k} \cdot \left[(U_2 - U_1) - V_2 \cdot z - V_3 \cdot z^2 \right]. \tag{15}$$

3.1.3. Measurement error. A central concern in the returns to education literature, also discussed by Card (2001), is measurement error in the level of schooling. Consider thus an extension of the previous model of the returns to education, where the true level of education X^* is observed only with measurement error, $X = X^* + V_4$. Suppose that Equations (13) and (15) hold with X^* taking the place of X:

$$Y^{x^*} = U_1 + U_2 \cdot x^* - \frac{k}{2} \cdot x^{*2}$$

$$X^{*z} = \frac{1}{k} \cdot \left[(U_2 - U_1) - V_2 \cdot z - V_3 \cdot z^2 \right].$$

Expressing both equations in terms of the noisy measurement X yields

$$Y^x = \left(U_1 - U_2 \cdot V_4 - \frac{k}{2} \cdot V_4^2 \right) + (U_2 + k \cdot V_4) \cdot X - \frac{k}{2} \cdot X^2 \tag{16}$$

$$X^z = \frac{1}{k} \cdot \left[(U_2 - U_1) - V_2 \cdot Z - V_3 \cdot Z^2 \right] + V_4. \tag{17}$$

This example illustrates that heterogeneity will in general be multi-dimensional in both equations of the triangular system if there is both true heterogeneity in terms of structural primitives as well as measurement error.

3.1.4. Partial market equilibrium. We conclude this section by discussing a special case of the systems of simultaneous equations analysed in Supplementary Appendix A. Consider partial equilibrium in some markets, where Y is the quantity sold in a given market, X is the price, and Z is some shifter of supply. Suppose that quantity demanded and quantity supplied at a given price x are given by

$$Y^x = U_1 + U_2 \cdot x + U_3 \cdot x^2 \tag{18}$$

$$Y_S^x = U_4 + U_5 \cdot x + U_6 \cdot x^2 + U_7 \cdot z, \tag{19}$$

where $U_1 - U_4 - U_7 \cdot z > 0$, $U_2 < 0$, $U_3 < 0$, $U_5 > 0$, $U_6 > 0$, and $U_7 > 0$ almost surely. The first condition guarantees positive excess demand when prices equal 0, the other conditions ensure monotonicity of demand and supply. Solving for equilibrium prices X^z equating demand Y^x and supply Y_S^x given the supply shifter $Z = z$, and requiring that X should be positive in equilibrium, yields the first stage relationship

$$X^z = \frac{U_5 - U_2 + \sqrt{(U_5 - U_2)^2 - 4 \cdot (U_3 - U_6) \cdot (U_1 - U_4 - U_7 \cdot z)}}{2 \cdot (U_1 - U_4 - U_7 \cdot z)}. \tag{20}$$

This example illustrates, in particular, that the reduced form of first-stage relationship corresponding to a system of simultaneous equations will, in general, not satisfy any separability which might allow to reduce the dimensionality of heterogeneity; this point was emphasized by Blundell and Matzkin (2010).

4. IDENTIFICATION

The following theorem presents our central identification result. It shows how to achieve identification of the distribution of the potential outcome Y^x (for a given x) in three steps.

(i) The distribution of Y^x given Z^x is identified from the distribution of Y given Z, X.
(ii) The distribution of Z^x is identified from the conditional distribution of X given Z.
(iii) Integrating $P(Y^x \leq y|Z^x)$ over the distribution of Z^x identifies the unconditional distribution of Y^x.

Point identification requires rich enough support of Z, *i.e.* the support of Z^x must be a subset of the support of Z.

Theorem 1. (Identification with a continuous instrument) *Under Assumptions 1, 2, 3, and 4*

$$P(Y^x \leq y|Z^x = z) = P(Y \leq y|X = x, Z = z) \tag{21}$$

for (x, z) *in the joint support of* (X, Z) *and*

$$F_{Z^x}(z) := P(Z^x \leq z) = P(X \geq x|Z = z) = 1 - F_{X|Z}(x|z). \tag{22}$$

for z *in the support of* Z. *Let* $p^x := P(Z^x \in [z_l, z_u]) = F_{Z^x}(z_u) - F_{Z^x}(z_l)$. *If* $p^x = 1$, *then*

$$P(Y^x \leq y) = \int_{z_l}^{z_u} P(Y \leq y|X = x, Z = z) dF_{Z^x}(z). \tag{23}$$

The proof of this theorem, and of all further results, can be found in Appendix A.

Remark Theorem 1 shows identification of the marginal distribution of Y^x. It does not show identification of the joint distribution of (Y^{x_1}, Y^{x_2}) for any x_1, x_2. In particular, it does not identify the distribution of the functions $g(., U)$. This is not surprising—even in the case of exogenous X (*i.e.* $X \perp U$) we can only identify the marginal distributions of Y^x if heterogeneity is left unrestricted.

Remark The identifiability shown in Theorem 1 contrasts interestingly with the arguments of Heckman and Vytlacil (2005). In the context of models with discrete X it is argued on p. 724 of Heckman and Vytlacil (2005) that "when we develop a symmetrically heterogeneous model, the method of instrumental variables breaks down entirely and a different approach to econometric policy analysis is required". Theorem 1 shows that for the case of continuous X this is not necessarily true.

4.1. *Identification using reweighting*

Theorem 1 shows that $P(Y^x \leq y)$ is identified from the distribution Y given X and Z, integrated over the distribution of Z^x. The following proposition restates this result in terms of the reweighted distribution of (Y, X, Z), using an identified weight function w.

Proposition 1. (Identification using reweighting) *Under Assumptions 1, 2, 3, and 4, suppose that the distribution of* Z^x *is absolutely continuous relative to the distribution of* Z *given* $X = x$. *Then we can define the weighting function*

$$w(x, z) := dF_{Z^x}(z)/dF_{Z|X}(z|x). \tag{24}$$

For this definition of w *we get*

$$P(Y^x \leq y) = E[\mathbf{1}(Y \leq y) \cdot w(X, Z)|X = x], \text{ and}$$
$$\overline{g}(x) = E[Y^x] = E[Y \cdot w(X, Z)|X = x]. \tag{25}$$

Remark If Z and Z^x are continuously distributed, we can differentiate their distribution functions, so that we can rewrite the weight function w as

$$w(x,z) = -\frac{\frac{\partial}{\partial z} F_{X|Z}(x|z)}{\frac{\partial}{\partial z} F_{Z|X}(z|x)}. \tag{26}$$

Proposition 1 shows that the weight w effectively "exogenizes" X, *i.e.* reweighting the distribution of (X,Y,Z) by $w(X,Z)$ implies that for the reweighted distribution $X \perp U$. To illustrate this, assume that all densities in the following display exist, the ratios involved are well defined and the support condition is fulfilled. Then Theorem 1 implies that

$$\int f_{U|X,Z}(u|x,z) f_{Z|X}(z|x) w(x,z) dz = \int f_{U|Z^x}(u|z) f_{Z^x}(z) dz = f_U(u), \tag{27}$$

so that after reweighting the distribution of U given $X = x$ is equal to the marginal distribution of U. We can thus run mean regressions or quantile regressions for the reweighted distribution and get unbiased estimates. Section 5 briefly discusses estimation based on this idea. An advantage of this estimation approach is its flexibility. It allows to estimate all kinds of objects of interest, such as QSFs or semi-parametric models, with the same weights, which only need to be constructed once.

4.2. *Partial identification*

The point identification result of Theorem 1 requires that the support of Z^x is contained in the support of Z. If this is not the case, the distribution of Y^x is still partially identified, as shown in the following proposition. The data point identify the distribution of Y^x given Z^x in the support of Z. The data are uninformative about the distribution of Y^x given Z^x for Z^x outside the support of Z.

Proposition 2. (Partial identification) *Under Assumptions 1, 2, 3, and 4, suppose that* $p^x = P(Z^x \in [z_l, z_u]) = F_{Z^x}(z_u) - F_{Z^x}(z_l) < 1$. *Then we can bound the distribution of* Y^x *by*

$$P(Y^x \leq y) \in \int_{z_l}^{z_u} P(Y \leq y | X = x, Z = z) dF_{Z^x}(z) + [0, 1 - p^x], \tag{28}$$

where these bounds are sharp.

Suppose additionally that Y has known bounded support $[\underline{y}, \overline{y}]$. *Then we can bound the ASF by*

$$\overline{g}(x) = E[Y^x] \in \int_{z_l}^{z_u} E[Y|X = x, Z = z] dF_{Z^x}(z) + [(1-p^x)\cdot\underline{y}, (1-p^x)\cdot\overline{y}], \tag{29}$$

where these bounds are sharp.

Note that the integral in equation (28) is bounded by p^x, so that the bound for $P(Y^x \leq y)$ never exceeds 1.

4.3. *The relationship to control functions*

Imbens and Newey (2009) propose to identify the ASF $\overline{g}(x)$ using a control function V^* which satisfies $X \perp U|V^*$. In particular, they consider

$$\tilde{g}(x) = \int_0^1 m(x,v)dv \tag{30}$$

where

$$m(x,v) := E[Y|X=x, V^*=v] \tag{31}$$

and $V^* = F_{X|Z}(X|Z)$. This definition of V^* implies that $V^* \sim U[0,1]$, so that equation (30) integrates over the marginal distribution of V^*. Imbens (2007) noted that conditional independence of X and U given V^* does not hold in the case of a random coefficient first stage. In Kasy (2011), it was shown more generally that conditional independence cannot be achieved if $\dim(V) > 1$.

The following theorem shows that the control function approach is nonetheless justified in the set-up considered in the present article, since it is equivalent to the reweighting given by equation (25) in Proposition 1.[5] The support requirement for point identification of $\tilde{g}(x)$ is that V^* has full support given $X=x$. This is equivalent to the requirement that $p^x = 1$ in Theorem 1.

Theorem 2. (Validity of the control function approach) *Under Assumptions 1, 2, 3, and 4, if $F_{X|Z}(X|Z)$ is differentiable in X and Z, and if $V^* = F_{X|Z}(X|Z)$ has full support given $X=x$, then*

$$P(Y^x \leq y) = \int_0^1 P(Y \leq y|X=x, V^*=v)dv, \tag{32}$$

even though in general

$$P(Y^x \leq y|V^*=v) \neq P(Y^x \leq y|X=x, V^*=v).$$

In particular, $\tilde{g}(x) = \overline{g}(x)$, where $\tilde{g}(x)$ is defined as in equation (30) *and $\overline{g}(x) = E[Y^x]$.*

Remark The motivation for control functions in Imbens and Newey (2009) is based on conditional independence $X \perp U|V^*$, which implies $P(Y^x \leq y|V^*=v) = P(Y^x \leq y|X=x, V^*=v)$. This in turn implies that equation (32) holds. In general, however, conditional independence cannot be achieved. Theorem 2 implies that we can nonetheless achieve independence of X and Y^x by conditioning on the control function V^* and averaging over its marginal distribution. Equation (32) does hold, even though its usual justification - conditional independence—does not hold. Supplementary Appendix D illustrates this point numerically in the context of a random coefficient model.

Remark Theorem 2 also generalizes to the case of less-than-full support of V^*. In that case, we get the same bounds on the ASF and QSF as derived in Proposition 2, and as discussed in Section 3.2 of Imbens and Newey (2009). Note, however, that the set-up of Imbens and Newey

5. In a working paper version of (Hahn and Ridder, 2011), a related result was shown which implies that the ASF is identified using the control function approach, as long as it is point identified. Their proof proceeds by constructing an observationally equivalent triangular system with one-dimensional first-stage heterogeneity, given any continuous distribution of (Y,X,Z).

(2009) also allows to talk about causal effects, ASFs, local policy effects, etc. for subpopulations defined in terms of V^*. This is not meaningful in our set-up, since the same value of V^* maps into different subpopulations in terms of V for different values of X.

Remark The proof of Theorem 2 proceeds by showing equivalence of the control function approach to the reweighting approach of Proposition 1. An alternative intuition for this result is as follows:[6] Define $V^{*x} := F_{X|Z}(x|Z^x)$. Then V^{*x} is a one-to-one transformation of Z^x, and by the same arguments as in the proof of Theorem 1 we get $P(Y \leq y|X=x, V^*=v) = P(Y^x \leq y|V^{*x}=v)$. V^{*x} furthermore has the property that it is uniformly distributed on $[0,1]$ for any x, since $V^{*x} = 1 - F_{Z^x}(Z^x)$, so that

$$\int_0^1 P(Y \leq y|X=x, V^*=v)dv = \int_0^1 P(Y^x \leq y|V^{*x}=v)dP_{V^{*x}}(v) = P(Y^x \leq y).$$

4.4. *Discrete instruments*

This section considers again the triangular system set-up of Section 3 with one modification. We drop the assumption of continuity of Z and assume instead that Z has finite support. It turns out that in this set-up the ASF is completely un-identified.

Theorem 3. (Non-identification with discrete instruments) *Suppose that Assumptions 1, 2, and 3 hold, that Z has finite support $z = 1, \ldots, \bar{z}$, and that X is continuously distributed with bounded density conditional on Z.*

Then the data are completely uninformative about the distribution of Y^x. That is, for any y the identified set for $P(Y^x \leq y)$, as a function of x, is given by $\{\tilde{g}^y : \tilde{g}^y$ is bounded by $[0,1]\}$.

Similarly, under the additional assumption that $Y \in [0,1]$, the identified set for the ASF $\overline{g}$ is given by $\{\overline{g} : \overline{g}$ is bounded by $[0,1]\}$. If we additionally impose continuity of g in x for all U, then the identified set is given by $\{\overline{g} : \overline{g}$ continuous and bounded by $[0,1]\}$.

Remark This theorem shows that, for the case of continuous treatment and discrete instrument, the data are completely uninformative about the distribution of Y^x, and in particular about the ASF. There is a "discontinuity" of identifiability, going from large but finite support of Z to a continuous support. This is counter-intuitive and owed to the fact that we have not precluded structural functions with arbitrarily high slopes or curvatures. An *a priori* bound on slopes or curvatures would lead to "continuity" of identifiability, but seems hard to justify in most cases. A more attractive alternative might be the imposition of a Bayesian prior which incorporates the assumption that high curvatures are unlikely, such as the Gaussian process priors discussed in Williams and Rasmussen (2006).

Remark It is interesting to compare the negative conclusion of Theorem 3 with the results of Torgovitsky (2011) as well as DíHaultfœuille and Février (2011), where it is shown that discrete instruments for a continuous treatment do have identifying power *if* the dimensionality of heterogeneity in both the first stage and the structural equation of interest is restricted. Note also that the identification problem implied by Theorem 3 results from discrete *instruments*, which is different from the problems arising for discrete *treatments* in the context of models with restricted heterogeneity, as discussed among others by Chesher (2005) and Shaikh and Vytlacil (2011).

6. I thank an anonymous referee for providing this argument.

5. ESTIMATION USING REWEIGHTING

Following Theorem 2, we can estimate the ASF by regressing Y on X and V^*, and averaging predicted values for $X=x$ over the marginal distribution of the control V^*. This is the approach proposed by Imbens and Newey (2009).

The identification result of Theorem 1 suggests an alternative estimation approach, based on reweighting. In particular, it was shown in Section 4.1 that, given sufficient support of the instrument, the ASF is identified by

$$\overline{g}(x)=E[Y\cdot w(X,Z)|X=x],$$

where

$$w(x,z)=\frac{f_{Z^x}(z)}{f_{Z|X}(z|x)}=-\frac{\frac{\partial}{\partial z}F_{X|Z}(x|z)}{\frac{\partial}{\partial z}F_{Z|X}(z|x)}.$$

Reweighting by w effectively makes X exogenous relative to the distribution of heterogeneity U. An estimator for w can be constructed as follows. Use a kernel density estimator for the denominator, *i.e.*

$$\widehat{f}(z|x)=\frac{1}{\tau_z}\frac{\sum_i K\left(\frac{Z_i-z}{\tau_z}\right)K\left(\frac{X_i-x}{\tau_x}\right)}{\sum_i K\left(\frac{X_i-x}{\tau_x}\right)}.$$

Use local linear regression, as well as smoothing of the indicator function, to estimate the numerator, *i.e.*

$$\widehat{f}_{Z^x}(z)=\operatorname*{argmin}_b \min_a \sum_i \left(L\left(\frac{x-X_i}{\tau_x}\right)-a-b\cdot(Z_i-z)\right)^2\cdot K\left(\frac{Z_i-z}{\tau_z}\right).$$

In these expressions, K is an appropriate kernel function integrating to one with support $[-1,1]$, L is the cumulative distribution function of a smooth symmetric distribution with support $[-1,1]$, and τ_x and τ_z are bandwidth parameters. Form an estimator for w from these two estimates,

$$\widehat{w}(x,z)=\frac{\widehat{f}_{Z^x}(z)}{\widehat{f}(z|x)}.$$

As previously discussed, these weights can be used to estimate best linear predictors, semi-parametric estimators, non-parametric estimators using series methods, etc. We will discuss non-parametric estimators using kernel regression.

Assuming that the support conditions for point identification are satisfied, we can estimate the ASF $E[Y^x]$ by the following reweighted kernel regression estimator,

$$\widehat{\beta}(x):=\sum_i Y_i\cdot K\left(\frac{X_i-x}{\tau_x}\right)\cdot\widehat{w}(X_i,Z_i)\Bigg/\sum_j K\left(\frac{X_j-x}{\tau_x}\right)\cdot\widehat{w}(X_j,Z_j). \tag{33}$$

This estimator is a special case of the partial means estimators considered by Newey (1994). All the asymptotic theory developed in Newey (1994) applies. The asymptotic variance of $c_n\cdot(\widehat{\beta}(x)-E(Y^x))$ (rescaled by an appropriate diverging sequence c_n), in particular, can be consistently estimated by c_n^2/n times the sample variance of the "influence function" of $\widehat{\beta}(x)$, so that

$$\operatorname{Var}\left(\widehat{\beta}(x)\right)\approx\frac{1}{n^2}\sum_i\widehat{\psi}_i^2,$$

where $\widehat{\psi}_i = \frac{\partial \widehat{\beta}(x)}{\partial p_n(y_i,x_i,z_i)}$. The derivative in the last expression is to be understood as the derivative of $\widehat{\beta}(x)$ with respect to the mass p_n put by the empirical distribution on the i-th observation. Details and background can be found in (van der Vaart, 2000, ch. 20) and Newey (1994).

Any application of the reweighting approach has to carefully consider support issues. In particular, if the ASF is only partially identified, there will be values of x,z for which the denominator of w equals 0. Even if the ASF is point identified, identification will in general be "weak" or "irregular", cf. Khan and Tamer (2010). This is a non-parametric analogue to the weak instrument problem, which has been discussed extensively in the context of linear models, see the review in Imbens and Wooldridge (2007). Weak identification is reflected in the unboundedness of $w(X,Z)$. Consistent estimators will, in general, have to apply trimming for observations with large estimated weights $\widehat{w}(x,z)$. As a consequence, the conditional expectation of weights given x will be <1, so that trimming involves a trade-off between bias and variance. The achievable rates of convergence depend on the tail distribution of $w(X,Z)$, and are not derived here.

Similar issues apply to the control function estimator of Imbens and Newey (2009). If we replace V^* and the conditional expectation $E[Y|X,V^*]$ with some (kernel) estimators, we get corresponding estimators for the ASF $E[Y^x]$ of the form

$$\begin{aligned}\widehat{\beta}^{CF}(x) &= \int_0^1 \widehat{E}[Y|X,\widehat{V}^*=v]dv \\ &= \sum_i Y_i \cdot \int_0^1 \left[\frac{K\left(\frac{X_i-x}{\tau_x}\right)\cdot K\left(\frac{\widehat{V}_i^*-v}{\tau_v}\right)}{\sum_j K\left(\frac{X_j-x}{\tau_x}\right)\cdot K\left(\frac{\widehat{V}_j^*-v}{\tau_v}\right)} \right] dv. \end{aligned} \tag{34}$$

This shows that the control function estimator can also be written as a reweighted average. If the density of V^* given x is not bounded away from 0, which is equivalent to the weight $w(x,z)$ being unbounded, the weighted average in equation (34) puts a lot of mass on some observations, which again leads to slower rates of convergence and the necessity of trimming.

6. EXTENSION TO HIGHER DIMENSIONS

We have so far assumed that the support of X and the support of Z are contained in $\mathbb{R}$. We will now discuss two extensions: (i) the overidentified case, where as before the support of X is one-dimensional, but now there are several continuously distributed instruments Z; and (ii) the case where both X and Z have k-dimensional support.

In the first case, we unsurprisingly find that the distribution of potential outcomes Y^x is still identified, and in fact overidentified, if the instruments have large enough support. In this case, it suffices to assume monotonicity of the first-stage relationship in one of the components of Z, without loss of generality Z_1. The (unconditional) distribution of Y^x is identified *conditional* on $Z_2 = (Z_{2,1},\ldots,Z_{2,k-1})$ as before, which immediately implies testable overidentifying restrictions and the possibility of constructing efficient estimators.

The case where X is of higher dimension turns out to be more complicated. In that case, the first part of our main identifying argument still generalizes and we get identification of the distribution of Y^x given Z^x, if the first stage is invertible in Z given V. As a consequence, we know that some weighting function $w = dP_{Z^x}/dP_{Z|X}$ exists which "exogenizes" X. There is no straightforward generalization of the assumption of monotonicity which would yield identification of the distribution of Z^x, and thus of w. If we are willing to impose additional exclusion restrictions

on the first-stage relationship, however, identification is possible. A triangular structure of h, in particular, allows to point-identify the distribution of Z^x.

6.1. *The overidentified case—one-dimensional X, multi-dimensional Z*

The following assumption replaces Assumptions 1 through 4. This assumption generalizes our basic set-up to the case of multi-dimensional Z, assuming monotonicity of the first-stage relationship in the first component of Z.

Assumption 5. (Overidentified triangular system)

1.

$$Y = g(X, U)$$
$$X = h(Z, V) \tag{35}$$

where X, Y, Z are random variables, X and Y take their values in $\mathbb{R}$ and Z takes its values in $\mathbb{R}^k$ ($k > 1$), the unobservables U, V have their support in an arbitrary measurable space of unrestricted dimensionality, and

$$Z \perp (U, V). \tag{36}$$

2. *Treatment X is continuously distributed in $\mathbb{R}$ conditional on Z.*
3. *The first-stage relationship $h(z, v)$ is strictly increasing in z_1 for all v and z_2, where $z = (z_1, z_2)$ and $z_1 \in \mathbb{R}$.*
4. *The instrument Z is continuously distributed in $\mathbb{R}^k$. The support of Z_1 given $Z_2 = z_2$ is $[z_{1l}(z_2), z_{1u}(z_2)]$. The first-stage relationship h is continuous in z_1 for all z and almost all v, and $P(X \leq x | Z = z)$ is continuous in z_1 for all x and z.*

Let, as before, $Y^x = g(x, U)$, $X^z = h(z, V)$, and

$$Z_1^x = \begin{cases} h^{-1}(x, Z_2, V) & \text{if } h(z_{1l}(Z_2), Z_2, V) \leq x \leq h(z_{1u}(Z_2), Z_2, V) \\ -\infty & \text{if } x < h(z_{1l}(Z_2), Z_2, V) \\ \infty & \text{if } h(z_{1u}(Z_2), Z_2, V) < x. \end{cases} \tag{37}$$

This definition exactly parallels the definition of Z^x, but conditions on Z_2 throughout. We get the following generalization of Theorem 1.

Proposition 3. (Identification with several instruments) *Suppose Assumption 5 holds. Then*

$$P(Y^x \leq y | Z_1^x = z_1, Z_2 = z_2) = P(Y \leq y | X = x, Z = (z_1, z_2)) \tag{38}$$

and

$$\begin{aligned} F_{Z_1^x | Z_2}(z_1 | z_2) &:= P(Z_1^x \leq z | Z_2 = z_2) \\ &= P(X \geq x | Z = (z_1, z_2)) = 1 - F_{X|Z}(x | z_1, z_2). \end{aligned} \tag{39}$$

Let

$$p^x(z_2) = F_{Z_1^x | Z_2}(z_1 | z_2)(z_{1u}(z_2)) - F_{Z_1^x | Z_2}(z_1 | z_2)(z_{1l}(z_2)).$$

If $p^x(z_2)=1$, *then*

$$P(Y^x \leq y) = \int_{z_{1l}(z_2)}^{z_{1u}(z_2)} P(Y \leq y | X=x, Z=(z_1,z_2)) dF_{Z_1^x|Z_2}(z_1|z_2). \tag{40}$$

Proposition 3 identifies $P(Y^x \leq y)$ using only the conditional distribution of (Y,X,Z_1) given $Z_2=z_2$. From this result, we can immediately derive overidentifying restrictions implied by the model, as well as alternative ways to estimate $P(Y^x \leq y)$, based on which we can derive efficient estimators. In the Supplementary Appendix C, some results are developed which provide estimators of minimal variance in the overidentified case.

6.2. *Multi-dimensional X and Z*

The following assumption again replaces Assumptions 1 through 4. It is a generalization of our basic set-up to the case of multi-dimensional X and Z, assuming invertibility of the first-stage relationship in Z given V.

Assumption 6. (Triangular system with multi-dimensional X and Z)

1. $$Y = g(X,U)$$
$$X = h(Z,V) \tag{41}$$
where X,Y,Z *are random variables,* Y *takes its values in* $\mathbb{R}$, *and* X *and* Z *take their values in* $\mathbb{R}^k$ $(k>1)$, *the unobservables* U,V *have their support in an arbitrary measurable space of unrestricted dimensionality, and*
$$Z \perp (U,V). \tag{42}$$
2. *Treatment* X *is continuously distributed in* $\mathbb{R}^k$ *conditional on* Z.
3. *The first-stage relationship* $h(z,v)$ *is invertible in* z *for all* v.
4. *The instrument* Z *is continuously distributed in* $\mathbb{R}^k$ *with support* $\mathscr{Z}$. *The first-stage relationship* h *is continuous in* z *for all* z *and almost all* v, *and* $P(X \leq x|Z=z)$ *is continuous in* z *for all* x.

In this setting, we can again define $Y^x = g(x,U)$, $X^z = h(z,V)$, and $Z^x = h^{-1}(x,V)$ if there is a $z \in \mathscr{Z}$ such that $h(z,V)=x$, and $Z^x=(\infty,\ldots,\infty)$ else. The first part of our identification argument immediately generalizes under this assumption. The second part of our identification argument generalizes if we impose additional exclusion restrictions. The following assumption provides an example of such exclusion restrictions.

Assumption 7. (Triangular first stage with multi-dimensional X and Z) *The first-stage relationship* h *has the following triangular structure:*

$$\begin{aligned} X_1 &= h_1(Z_1,V) \\ X_2 &= h_2(Z_1,Z_2,V) \\ &\vdots \\ X_k &= h_k(Z_1,\ldots,Z_k,V), \end{aligned} \tag{43}$$

where h_j is strictly increasing in z_j for all z, v, and j.

Theorem 4. *Suppose Assumption 6 holds. Then*

$$P(Y^x \leq y|Z^x = z) = P(Y \leq y|X = x, Z = z) \tag{44}$$

and, if $P(Z^x \in \mathscr{Z}) = 1$,

$$P(Y^x \leq y) = \int P(Y \leq y|X = x, Z = z) dP_{Z^x}(z). \tag{45}$$

If additionally the distribution of Z^x is absolutely continuous relative to the distribution of Z given $X = x$, then

$$\begin{aligned} P(Y^x \leq y) &= E[\mathbf{1}(Y \leq y) \cdot w(X,Z)|X = x], \text{ and} \\ \overline{g}(x) &= E[Y^x] = E[Y \cdot w(X,Z)|X = x]. \end{aligned} \tag{46}$$

for the weighting function[7]

$$w(x,z) = dP_{Z^x}(z)/dP_{Z|X}(z|x). \tag{47}$$

Suppose now that Assumptions 6 and 7 hold.[8] *Then Z_j^x does not depend on $x_{j+1}, \ldots, x_k$, and the distribution of Z^x is identified by $P(Z_1^x \leq z_1) = P(X_1 \geq x_1|Z_1 = z_1)$ and*

$$\begin{aligned} &P(Z_j^x \leq z_j|Z_1^x = z_1, \ldots, Z_{j-1}^x = z_{j-1}) \\ &= P(X_j \geq x_j|X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, Z_1 = z_1, \ldots, Z_j = z_j) \end{aligned} \tag{48}$$

for all $j = 2, \ldots, k$.

Remark

- The second part of Theorem 1, which allows to identify P_{Z^x}, unfortunately does not generalize to higher dimensions without imposing additional exclusion restrictions, as in Assumption 7. The problem, which is in some sense a non-parametric "reverse regression" problem, can be described as follows. The underlying structure of the first-stage relationship $X^z = h(z, V)$ can be equivalently described in terms of a joint distribution of X^z across values of z. The data identify only the marginal distribution of X^z for values of z in the support of Z. What we need are the (marginal) distributions of $Z^x = h^{-1}(x, V)$. These marginal distributions are identified under a given set of assumptions if and only if the mapping from the joint distribution of the $\{X^z\}_{z \in \mathscr{Z}}$ to the distribution of Z^x depends only on the marginal distributions of X^z. As was shown in Theorem 1, this is the case when X and Z are one-dimensional, and h is assumed to be strictly monotonic.
- To see that there is no straightforward generalization of the assumption of monotonicity to the multi-dimensional case, consider the case where $\dim(X) = \dim(Z) > 1$. Assume that

7. The expression $dP_{Z^x}(z)/dP_{Z|X}(z|x)$ here is used to denote the Radon–Nikodym derivative, or relative density, of Z^x relative to Z given X. This relative density exists by the assumption of absolute continuity.

8. I owe a variant of the following result to an anonymous referee.

the function h is differentiable and strictly monotonic componentwise in the sense that

$$s_{ij} := \operatorname{sign}\left(\frac{\partial}{\partial z_j} h_i(z, U)\right) \neq 0$$

is constant in z and U for all i, j. This assumption immediately implies

$$(X_i \leq x_i) \Rightarrow s_{ij} \cdot (Z_j^x - Z_j) \geq 0 \text{ for some } j \tag{49}$$

and

$$(X_i \geq x_i) \Rightarrow s_{ij} \cdot (Z_j^x - Z_j) \leq 0 \text{ for some } j, \tag{50}$$

for all i. Every component of X does, therefore, allow to exclude one of two opposite orthants for the difference $Z^x - Z$. If the rows of the matrix $S = (s_{ij})$ are linearly independent, we can therefore exclude k orthants. In total there are 2^k orthants. In the one-dimensional case, we can therefore pin down in which orthant (halfline) $Z^x - Z$ lies, and therefore—using independence—pin down $P(Z^x - z > 0)$. If k is >1, however, this is not possible anymore. For $k = 2$, for instance, these inequalities only pin down a half-plane for $Z^x - Z$, which is not sufficient in order to identify the distribution of Z^x.

7. CONCLUSION

This article discusses identification in continuous triangular systems without any restrictions on heterogeneity or functional form. We do assume monotonicity of the first-stage relationship in the instrument. We show that, under this condition alone, the ASF is identified if the instrument has rich enough support. Furthermore, the control function approach of Imbens and Newey (2009) does identify the ASF despite the fact that conditional independence does not hold if heterogeneity is multi-dimensional. We show that the ASF is partially identified if the instrument is continuous with insufficient support. The ASF remains completely unidentified if the instrument is discrete. These results immediately apply to simultaneous systems with unrestricted heterogeneity, since such systems give rise to a reduced form satisfying our assumptions.

APPENDIX A: PROOFS

Proof of Theorem 1. Equation (21) holds because

$$\begin{aligned} &P(Y \leq y | X = x, Z = z) \\ &= P(Y^x \leq y | X = x, Z = z) \\ &= P(Y^x \leq y | Z^x = z, Z = z) \\ &= P(Y^x \leq y | Z^x = z). \end{aligned} \tag{A.1}$$

The first equality follows from the definition of Y^x. The second equality follows because $X = x$ and $Z = z$ holds if and only if $Z^x = z$ and $Z = z$. The last equality follows from exogeneity of Z; $(U, V) \perp Z$ implies $(Y^x, Z^x) \perp Z$.

To see that equation (22) holds, note that the monotonicity Assumption 3 implies that $Z^x \leq z$ if and only if $X^z \geq x$. Exogeneity of the instrument implies

$$P(X^z \geq x) = P(X \geq x | Z = z) = 1 - P(X \leq x | Z = z),$$

where the last equality follows from continuity of the distribution of X given Z.

Equation (23) follows if $p^x = 1$, since

$$P(Y^x \leq y) = \int_{z_l}^{z_u} P(Y^x \leq y | Z^x = z) dF_{Z^x}(z) = \int_{z_l}^{z_u} P(Y \leq y | X = x, Z = z) dF_{Z^x}(z). \qquad \|$$

Proof of Proposition 1. To show the reweighting representation of the ASF given in equation (25), note that

$$\begin{aligned}\overline{g}(x) &= \int E[Y|X=x,Z=z]dF_{Z^x}(z)\\ &= \int E[Y|X=x,Z=z]\frac{dF_{Z^x}(z)}{dF_{Z|X}(z|x)}dF_{Z|X}(z|x)\\ &= \int E\left[Y\cdot\frac{dF_{Z^x}(z)}{dF_{Z|X}(z|x)}\middle|X=x,Z=z\right]dF_{Z|X}(z|x)\\ &= E[Y\cdot w(X,Z)|X=x]. \end{aligned} \tag{A.2}$$

The reweighting representation for $P(Y^x\leq y)$ follows by the same argument, once we replace Y by $\mathbf{1}(Y\leq y)$. ‖

Proof of Proposition 2. To derive the bounds on $P(Y^x\leq x)$ for the case that $p^x<1$, note that

$$P(Y^x\leq y)=P(Y^x\leq x|Z^x\in[z_l,z_u])\cdot p^x+P(Y^x\leq x|Z^x\notin[z_l,z_u])\cdot(1-p^x),$$

where

$$P(Y^x\leq y|Z^x\in[z_l,z_u])=\frac{1}{p^x}\int_{z_l}^{z_u}P(Y\leq y|X=x,Z=z)dF_{Z^x}(z).$$

Under our assumptions, the data impose no restrictions on $P(Y^x\leq x|Z^x\notin[z_l,z_u])$, which implies sharpness of these bounds.

The bounds on $\overline{g}$ follow by exactly the same argument. ‖

Proof of Theorem 2. By similar arguments as in the derivation of the reweighting result of Theorem 1 (note that $dF_{V^*}\equiv 1$),

$$\begin{aligned}&\int_0^1 P(Y\leq y|X=x,V^*=v)dv\\ &= \int E[\mathbf{1}(Y\leq y)|X=x,V^*]\frac{dF_{V^*}(V^*)}{dF_{V^*|X}(V^*|x)}dF_{V^*|X}(V^*|x)\\ &= \int E\left[\mathbf{1}(Y\leq y)\cdot\frac{dF_{V^*}(V^*)}{dF_{V^*|X}(V^*|x)}\middle|X=x,V^*=v\right]dF_{V^*|X}(V^*|x)\\ &= E[\mathbf{1}(Y\leq y)\cdot\tilde{w}(X,Z)|X=x], \end{aligned} \tag{A.3}$$

where

$$\tilde{w}(X,Z)=\frac{dF_{V^*}(V^*)}{dF_{V^*|X}(V^*|X)}=\frac{1}{f_{V^*}(V^*|X)}.$$

The weight $\tilde{w}$ is a function of (X,Z), since V^* is a function of (X,Z). To calculate this weight we need the conditional density of $V^*=F_{X|Z}(X|Z)$ given X. Recall that Assumption 3 implies that the mapping from Z to V^* given X is invertible. The conditional density of V^* given X thus equals

$$f_{V^*|X}(v|x)=f_{Z|X}(z|x)/\left|\frac{\partial v}{\partial z}\right|=\frac{dF_{Z|X}(z|x)}{-\frac{\partial}{\partial z}F_{X|Z}(x|z)}$$

by a change of variables, where z in these expressions is such that $F_{X|Z}(x|z)=v$. It follows that $\tilde{w}(X,Z)=w(X,Z)$. This implies that the control function approach is equivalent to the reweighting approach in Theorem 1.

The result that (in general) $P(Y^x\leq y|V^*=v)\neq P(Y^x\leq y|X=x,V^*=v)$ follows from the arguments in Kasy (2011), and is also confirmed by the example discussed in Supplementary Appendix D. ‖

Proof of Theorem 3. We will focus on the identified set for $\overline{g}(x)=E[Y^x]$, under the assumption of bounded Y. The claim for $P(Y^x\leq y)$ follows by the same argument. We show, first, that $\overline{g}\equiv 0$ is in the closure of the identified set for $\overline{g}$, before considering the general claim..

The proof is constructive, providing an explicit distribution of structural functions $P(g(.,U))$ consistent with the data. The proof has two steps: (i) first we need to construct a coupling of the marginal distributions $P((X,Y)|Z=z)=P(X^z,Y^z)=P(h(z,V),g(h(z,V),U))$. Coupled values of (X,Y) for different values of Z are those assumed to have the same realization of (U,V). (ii) In the second step we show, by filling in intermediate values of $g(.,U)$ for values of X that are not realized for any Z, that any *ASF* satisfying the *a priori* bounds can be achieved.

(i) The data and assumptions identify the (marginal) distributions of (X^z,Y^z) for $z=1\ldots\bar{z}$, since $P((Y^z,X^z))=P((Y,X)|Z=z)$. The coupling of these marginal distributions is not identified, but restricted by monotonicity of the first stage. Any joint distribution of $(X^z,Y^z)_{z=1\ldots\bar{z}}$ which respects the monotonicity of the first stage is consistent with the

observed data distribution and the model assumptions. We can choose for instance the "rank-invariance" coupling, which satisfies $F_{X|Z}(X^z|z)=F_{X|Z}(X^{z'}|z')$,[9] and $F_{Y|X,Z}(Y^z|X^z,z)=F_{Y|X,Z}(Y^{z'}|X^{z'},z')$.

(ii) Now take any realized value from this joint distribution of $\{(X^z,Y^z):z=1,\ldots,\bar{z}\}$. For this realized value construct $g(.,U)$ as follows, assuming that Y has continuous support and g is required to be continuous:[10]

$$g(x,U)=Y^x=\sum_{z=1}^{\bar{z}}K\left(\frac{x-X^z}{\tau'}\right)Y^z,$$

where K is a continuous and positive kernel function with support $[-1,1]$ and $K(0)=1$ such that $K(x)$ is increasing for $x\leq 0$ and decreasing for $x\geq 0$, and τ' is a (random) "bandwidth" parameter, given by

$$\tau'=\min(\tau,\min_{z,z'}(|X^z-X^{z'}|)).$$

τ' is positive with probability 1 by continuity of the distribution of X^z. With this construction we get

$$\overline{g}(x)=E[Y^x]=\sum_{z=1}^{\bar{z}}E\left[K\left(\frac{x-X^z}{\tau'}\right)Y^z\right].$$

Since we assumed that X is continuously distributed with bounded density conditional on Z, choosing τ small enough we can make this expression arbitrarily small, uniformly in x, as

$$\begin{aligned}\left|E\left[K\left(\frac{x-X^z}{\tau'}\right)Y^z\right]\right|&\leq E\left[K\left(\frac{x-X^z}{\tau'}\right)\right]\\&\leq E\left[K\left(\frac{x-X^z}{\tau}\right)\right]\\&\leq\tau\int_{-\infty}^{\infty}K(\tilde{x})d\tilde{x}\cdot\sup_x f_{X^z}(x)=O(\tau).\end{aligned}$$

The first inequality holds since $Y\in[0,1]$. The second inequality holds since by definition of τ' we have $\tau'\leq\tau$, and by the two-sided monotonicity of K. The supremum $\sup_x f_{X^z}(x)$ is finite by assumption. This proves that $\overline{g}\equiv 0$ is in the closure of the identified set for $\overline{g}$.

To see that the same holds for any function $\tilde{g}$, consider the same construction, but replace $g(x,U)$ by

$$g(x,U)=Y^x=\sum_{z=1}^{\bar{z}}\left[K\left(\frac{x-X^z}{\tau'}\right)Y^z+\left(1-K\left(\frac{x-X^z}{\tau'}\right)\tilde{g}(x)\right)\right].$$ ‖

Proof of Proposition 3. This is an immediate consequence of Theorem 1, once we note that Assumptions 1 through 4 hold *conditional* on Z_2. ‖

Proof of Theorem 4. As before, the first part of this theorem follows from

$$\begin{aligned}&P(Y\leq y|X=x,Z=z)\\&=P(Y^x\leq y|X=x,Z=z)\\&=P(Y^x\leq y|Z^x=z,Z=z)\\&=P(Y^x\leq y|Z^x=z).\end{aligned}\tag{A.4}$$

Under Assumption 7, the exclusion restrictions imposed by the triangular structure immediately imply that Z_j^x is a function of $x_1,\ldots,x_j$ only, and that X_j^z is a function of $z_1,\ldots,z_j$ only. The exclusion restrictions in combination with the monotonicity conditions imply furthermore that $Z_j^x\leq z_j$ and $Z_1^x=z_1,\ldots,Z_{j-1}^x=z_{j-1}$ if and only if $X_j^z\geq x_j$ and $Z_1^x=z_1,\ldots,Z_{j-1}^x=z_{j-1}$.

9. This is the coupling consistent with the assumption of monotonicity of h in a one-dimensional unobservable V. The proof of this theorem thus holds verbatim when h is assumed to be monotonic in V.

10. If g is not required to be continuous, and in particular if Y has discrete support, then we can simply set $g(x,U)=0$ for $x\notin\{X^1,\ldots,X^{\bar{z}}\}$, and the remaining argument becomes trivial. This is, in particular, the case for identification of $P(Y^x\leq y)$, where $\mathbf{1}(Y\leq y)$ takes the place of Y.

Identification of $P(Z_1^x \leq z_1)$ follows as in the univariate case. Identification of $P(Z_j^x \leq z_j | Z_1^x = z_1, \ldots, Z_{j-1}^x = z_{j-1})$ follows from

$$\begin{aligned}
&P(X_j \geq x_j | X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, Z_1 = z_1, \ldots, Z_j = z_j) \\
=&P(X_j^z \geq x_j | X_1^z = x_1, \ldots, X_{j-1}^z = x_{j-1}, Z_1 = z_1, \ldots, Z_j = z_j) \\
=&P(X_j^z \geq x_j | Z_1^x = z_1, \ldots, Z_{j-1}^x = z_{j-1}, Z_1 = z_1, \ldots, Z_j = z_j) \\
=&P(Z_j^x \leq z_j | Z_1^x = z_1, \ldots, Z_{j-1}^x = z_{j-1}).
\end{aligned}$$

‖

Acknowledgments. I would like to thank Stephane Bonhomme, Gary Chamberlain, Bryan Graham, Jinyong Hahn, Stefan Hoderlein, Guido Imbens, Susanne Kimm, Zhipeng Liao, Rosa Matzkin, James Powell, Geert Ridder, Alexander Rothenberg, James Stock, as well as several anonymous referees for valuable discussions and comments.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

BLUNDELL, R. and MATZKIN, R. (2010), "Conditions for the Existence of Control Functions in Nonseparable Simultaneous Equations Models" (Unpublished manuscript).
BLUNDELL, R. and POWELL, J. (2003), "Endogeneity in Non-parametric and Semi-parametric Regression Models", in *Advances in Economics and Econometrics: Theory and Applications, Eighth world Congress*, vol. 2, pp. 655–679.
CARD, D. (2001), "Estimating the Return to Schooling: Progress on some Persistent Econometric Problems", *Econometrica*, **69**, 1127–1160.
CHERNOZHUKOV, V. and HANSEN, C. (2005), "An IV Model of Quantile Treatment Effects", *Econometrica*, **73**, 245–261.
CHERNOZHUKOV, V., IMBENS, G. W. and NEWEY, W. K. (2007), "Instrumental Variable Estimation of Nonseparable Models", *Journal of Econometrics*, **139**, 4–14.
CHESHER, A. (2003), "Identification in Nonseparable Models", *Econometrica*, **71**, 1405–1441.
CHESHER, A. (2005), "Non-parametric Identification Under Discrete Variation", *Econometrica*, **73**, 1525–1550.
DE CHAISEMARTIN, C. (2012), "All You Need is LATE" (Working Paper, Warwick University).
DÍHAULTFOEUILLE, X. and FÉVRIER, P. (2011), "Identification of Nonseparable Models with Endogeneity and Discrete Instruments" (Working Paper, CREST).
FLORENS, J. P., HECKMAN, J. J., MEGHIR, C. and VYTLACIL, E. (2008), "Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects", *Econometrica*, **76**, 1191–1206.
GAUTIER, E. and HODERLEIN, S. (2011), "A Triangular Treatment Effect Model with Random Coefficients in the Selection Equation", *arXiv preprint arXiv:1109.0362*.
HAHN, J. and RIDDER, G. (2011), "Conditional Moment Restrictions and Triangular Simultaneous Equations", *The Review of Economics and Statistics*, **93**, 683–689.
HAN, S. (2012), "Non-parametric Triangular Simultaneous Equations Models with Weak Instruments" (Working Paper, UT Austin).
HECKMAN, J. and VYTLACIL, E. (2005), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation", *Econometrica*, **73**, 669–738.
HECKMAN, J. J., SMITH, J. and CLEMENTS, N. (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.
HODERLEIN, S., KLEMELÄ, J. and MAMMEN, E. (2010), "Analyzing the Random Coefficient Model Non-Parametrically", *Econometric Theory*, **26**, 804–837.
HODERLEIN, S. and MAMMEN, E. (2007), "Identification of Marginal Effects in Nonseparable Models Without Monotonicity", *Econometrica*, **75**, 1513–1518.
HODERLEIN, S. and SASAKI, Y. (2013), "Outcome Conditioned Treatment Effects", Technical Report, (Discussion Paper, Boston College & Johns Hopkins University).
HOROWITZ, J. L. (2011), "Applied Non-parametric Instrumental Variables Estimation", *Econometrica*, **79**, 347–394.
HUBER, M. and MELLACE, G. (2010), "Sharp Bounds on Average Treatment Effects under Sample Selection" (Working Paper, University of St Gallen).
IMBENS, G. (2007), "Nonadditive Models with Endogenous Regressors", *Econometric Society World Congress Series*, **75**, 17–46.
IMBENS, G. W. and ANGRIST, J. (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, **62**, 467–467.

IMBENS, G. W. and NEWEY, W. (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", *Econometrica*, **77**, 1481–1512.
IMBENS, G. W. and WOOLDRIDGE, J. M. (2007), "What's New in Econometrics? Weak instruments and many instruments", *NBER Lecture Notes 13, Summer 2007*.
JUN, S. J., PINKSE, J. and XU, H. (2011), "Tighter Bounds in Triangular Systems", *Journal of Econometrics*, **161**, 122–128.
JUN, S. J., PINKSE, J., XU, H. and YILDIZ, N. (2010), "Identification of Treatment Effects in a Triangular System of Equations. Technical report, (Discussion Paper 130910, UT Austin).
KASY, M. (2011), "Identification in Triangular Systems using Control Functions", *Econometric Theory*, **27**, 663–671.
KHAN, S. and TAMER, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation", *Econometrica*, **78**, 2021–2042.
MANSKI, C. F. (2003) *Partial Identification of Probability Distributions* (New York: Springer).
MANSKI, C. F. and PEPPER, J. V. (2003), "Monotone Instrumental Variables: With an Application to the Returns to Schooling", *Econometrica*, **68**, 997–1010.
MATZKIN, R. (2008), "Identification in Non-parametric Simultaneous Equations Models", *Econometrica*, **76**, 945–978.
NEWEY, W., POWELL, J. and VELLA, F. (1999), "Non-parametric Estimation of Triangular Simultaneous Equations Models", *Econometrica*, **67**, 565–603.
NEWEY, W. K. (1994), "Kernel Estimation of Partial Means and a General Variance Estimator", *Econometric Theory*, **10**, 233–253.
NEWEY, W. K. and POWELL, J. L. (2003), "Instrumental Variable Estimation of Non-parametric Models", *Econometrica*, **71**, 1565–1578.
SCHENNACH, S., WHITE, H. and CHALAK, K. (2012), "Local Indirect Least Squares and Average Marginal Effects in Nonseparable Structural Systems", *Journal of Econometrics*, **166**, 282–302.
SHAIKH, A. M. and VYTLACIL, E. J. (2011), "Partial Identification in Triangular Systems of Equations with Binary Dependent Variables", *Econometrica*, **79**, 949–955.
TORGOVITSKY, A. (2011), "Identification of Nonseparable Models with General Instruments" (Working Paper, Northwestern University).
VAN DER VAART, A. (2000), *Asymptotic Statistics* (Cambridge: Cambridge University Press).
WILLIAMS, C. and RASMUSSEN, C. (2006), *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press).