Econometric Theory, **27**, 2011, 663–671.
doi:10.1017/S0266466610000460

NOTES AND PROBLEMS

IDENTIFICATION IN TRIANGULAR SYSTEMS USING CONTROL FUNCTIONS

MAXIMILIAN KASY
University of California, Berkeley

This note discusses identification in nonparametric, continuous triangular systems. It provides conditions that are both necessary and sufficient for the existence of control functions satisfying conditional independence and support requirements. Confirming a commonly noticed pattern, these conditions restrict the admissible dimensionality of unobserved heterogeneity in the first-stage structural relation, or more generally the dimensionality of the family of conditional distributions of second-stage heterogeneity given explanatory variables and instruments. These conditions imply that no such control function exists without assumptions that seem hard to justify in most applications. In particular, none exists in the context of a generic random coefficient model.

1. INTRODUCTION

In a recent paper, Imbens and Newey (2009) develop nonparametric identification results in triangular systems for models with a first stage that is monotonic in unobservables. Based on these results, they construct inference procedures. In related work, Imbens (2007) surveys control functions in triangular systems more generally. This note elaborates on their analysis by providing conditions that are both necessary and sufficient for the existence of control functions. The nonparametric, continuous triangular system setup considered is given by

$$Y = g(X, \epsilon), \tag{1}$$

$$X = h(Z, \eta), \tag{2}$$

where we assume

$$Z \perp (\epsilon, \eta), \tag{3}$$

I thank Giuseppe Cavaliere, Bryan Graham, James Powell, Sarath Sanga, and two anonymous referees for valuable discussions and comments. Address correspondence to Maximilian Kasy, Department of Economics, UC Berkeley, 508-1 Evans Hall 3880, Berkeley, CA 94720-3880; e-mail: maxkasy@econ.berkeley.edu.

© Cambridge University Press 2010

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

with Z, X, Y each continuously distributed in $\mathbb{R}$. This is the general setup in which the application of instrumental variable methods is usually discussed, where Z is the exogenous instrument, X is the treatment, and Y is the outcome variable. The object of interest is the structural function g.

Recent contributions to the literature have generalized identification in parametric (linear) triangular models to nonparametric setups. The idea of nonparametric identification using the control function approach is to find a function C of X and Z such that, for $V = C(X, Z)$,

$$X \perp \epsilon | V. \tag{4}$$

If C is a one-dimensional, strictly monotonic function of both X and Z, then there exists a one-to-one mapping between (X, Z), (X, V), and (Z, V). Existence of an invertible mapping between (X, V) and (Z, V) implies that conditional independence (4) is equivalent to

$$Z \perp \epsilon | V. \tag{5}$$

In contrast to the literature providing mostly sufficient conditions for identification, this note provides conditions that are both necessary and sufficient for the existence of control functions that satisfy conditional independence and support requirements. These conditions impose restrictions on the dimensionality of unobserved heterogeneity. While the importance of dimensionality restrictions has been noted repeatedly, among others by Imbens (2007), this note is, to the best of our knowledge, the first to formally show that they are both necessary and sufficient.

The central object of interest in the control function literature is the average structural function (ASF). Let E_ϵ denote the expectation taken over the marginal distribution of ϵ, and similarly for E_V. The ASF was defined by Blundell and Powell (2003) as $ASF(x) := E_\epsilon[g(x, \epsilon)]$. Given a control function, the ASF is identified by

$$ASF(x) = \mathrm{E}_V[\mathrm{E}[g(X, \epsilon)|V, X = x]] = \mathrm{E}_V[\mathrm{E}[Y|V, X = x]]. \tag{6}$$

The first equality requires conditional independence (4). Identification of the conditional expectation given X, V requires full support of V given X. Under the same conditions, one can identify the quantile structural function (QSF) $g^\tau(x)$, which is given by the τth quantile of $g(x, \epsilon)$ over the marginal distribution of ϵ, as well as functions defined by more general conditional moment restrictions.[1]

Various choices for C have been proposed in the literature. Newey, Powell, and Vella (1999) suggest using the residual of a first-stage mean regression:

$$V = C(X, Z) = X - \mathrm{E}[X|Z]. \tag{7}$$

This is justified by an additive model for h, i.e., $h(Z, \eta) = \tilde{h}(Z) + \eta$ and $V = \eta$. Such additivity would not hold, for example, in models with heteroskedastic residuals.

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

Imbens and Newey (2009) propose, more generally, to use the conditional cumulative distribution function F of X given Z:

$$V = C(X, Z) = F[X|Z]. \tag{8}$$

This is justified by a first-stage h that is strictly monotonic in a one-dimensional η, implying that $V = F(\eta)$.

In either of these cases, the fact that V is a function of η alone immediately implies that conditional independence (5) holds. Under strict monotonicity of the conditional expectation or distribution function in Z, this in turn implies conditional independence (4).

The next section provides an example of failure of conditional independence (4) using the control function (8). Sections 3.1 and 3.2 provide the central results of this note. Section 3.1 gives a condition that is both necessary and sufficient for the existence of a control function that is constant in the instrument given unobserved heterogeneity. Section 3.2 does the same for the more general case of control functions satisfying conditional independence (4). Section 3.2 also shows that no control function can exist in the case of the random coefficient model. Section 3.3 states extensions to the case of higher dimensional X and Z. Section 4 concludes. All proofs are relegated to the Appendix.

2. COUNTEREXAMPLE

Consider the following random coefficient model:

$$X = \eta_1 + \eta_2 Z = \eta \cdot (1, Z), \tag{9}$$

$$(\eta_1, \eta_2, \epsilon) \sim N(\mu, \Sigma), \tag{10}$$

$$Z \perp (\eta, \epsilon). \tag{11}$$

This model will serve as a counterexample for identification attempts using control functions. Imbens (2007, Sect. 5.2) uses the same example of the failure of the control function of Imbens and Newey (2009). As will be shown in Section 3, no control function exists in this specification because first-stage heterogeneity η is more than one-dimensional.

For an economic example of this,[2] suppose the following: We are interested in the production function relating output of firms to a single variable input, e.g., labor l. Production technology is Cobb-Douglas, i.e., log output is $y_i = A_i + \alpha_i l_i$, where A and α are unobserved heterogeneity in firm technology or endowment with other factors. Prices for the output good vary exogenously, wages are constant, and firms maximize profits. Then both the first-stage relationship, that is, firm-specific labor demand as a function of prices, and the second-stage production function exhibit a linear random coefficient structure.

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

A similar example is the problem of estimating returns to schooling when returns are heterogeneous, schooling depends on returns, and we observe an independent cost variable affecting school choice that can serve as an instrument.

Now we will show why the control function proposed by Imbens and Newey (2009), $V = F(X|Z)$, fails in this random coefficient model. For jointly normally distributed variables, the conditional expectation is given by the best linear predictor. Hence we get, by ordinary least squares regression of ϵ on X given Z,

$$\mathrm{E}[\epsilon|X,Z] = \mu_\epsilon + (X - \mathrm{E}[X|Z]) \cdot \frac{\mathrm{Cov}(X,\epsilon|Z)}{\mathrm{Var}(X|Z)}.$$

The assumptions imply that X and ϵ are jointly normal given Z, with

$$\mathrm{Cov}(X,\epsilon|Z) = \Sigma_{\eta_1,\epsilon} + Z\Sigma_{\eta_2,\epsilon},$$
$$\mathrm{Var}(X|Z) = \Sigma_{\eta_1,\eta_1} + 2Z\Sigma_{\eta_1,\eta_2} + Z^2\Sigma_{\eta_2,\eta_2},$$
$$\mathrm{E}[X|Z] = \mu_{\eta_1} + Z\mu_{\eta_2}.$$

This gives

$$\mathrm{E}[\epsilon|X,Z] = \mu_\epsilon + (X - \mu_{\eta_1} - Z\mu_{\eta_2}) \cdot \frac{\Sigma_{\eta_1,\epsilon} + Z\Sigma_{\eta_2,\epsilon}}{\Sigma_{\eta_1,\eta_1} + 2Z\Sigma_{\eta_1,\eta_2} + Z^2\Sigma_{\eta_2,\eta_2}}. \tag{12}$$

The control function proposed by Imbens and Newey (2009),

$$V = F(X|Z) = \Phi\left(\frac{(X - \mu_{\eta_1} - Z\mu_{\eta_2})}{\sqrt{\mathrm{Var}(X|Z)}}\right), \tag{13}$$

is monotonic in X. If the support of Z is restricted to an appropriate range,[3] it is also monotonic in Z. Hence, for at least a subrange of V, the following equalities hold:

$$\begin{aligned}\mathrm{E}[\epsilon|V,X] &= \mathrm{E}[\epsilon|V,Z] = \mathrm{E}[\epsilon|X,Z] \\ &= \mu_\epsilon + \Phi^{-1}(V)\frac{\Sigma_{\eta_1,\epsilon} + Z\Sigma_{\eta_2,\epsilon}}{\sqrt{\Sigma_{\eta_1,\eta_1} + 2Z\Sigma_{\eta_1,\eta_2} + Z^2\Sigma_{\eta_2,\eta_2}}}.\end{aligned} \tag{14}$$

From this it follows that conditional independence (4) is violated: By invertibility of C in both X and Z, conditional independence (4) and (5) are equivalent. Conditional independence (5) requires conditional mean independence, i.e., that $\mathrm{E}[\epsilon|V,Z]$ is constant in Z given V. By equation (14) this holds if and only if $\Sigma_{\eta_2,\eta_2} = 0$, that is, if the slope of the first stage is constant. If the slope η_2 has positive variance, conditional independence (4) does not hold. A similar argument can be made about the conditional variance of ϵ given V, X.

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

3. CHARACTERIZATION OF MODELS IN WHICH CONTROL FUNCTIONS EXIST

The next three subsections will present the general results characterizing triangular systems for which control functions exist. Section 3.1 shows that control functions that do not depend on Z given η exist if and only if η is one-dimensional. This requirement is in particular violated by the random coefficient model of the previous section. Section 3.2 shows that control functions that satisfy conditional independence (4) exist if and only if the family of conditional distributions $P(\epsilon|X, Z)$ is one-dimensional. This dimensionality requirement is again violated by the random coefficient model. Section 3.3 finally generalizes the previous results to setups with higher-dimensional X and Z.

3.1. Control Functions that Do Not Depend on Z Given η

The next proposition covers all variants of the control function approach that we are aware of, in particular Newey, et al. (1999) and Imbens and Newey (2009).

PROPOSITION 1. *If $V = C(h(Z, \eta), Z)$ does not depend on Z given η, then conditional independence (5) holds.*

As mentioned in the Introduction, conditional independence (5) is equivalent to (4) if there exists a mapping $(Z, V) \to (X, V)$, which is true if C is invertible. Conditional independence (4) is necessary for the use of V as a control. The condition of Proposition 1, however, comes at the price of restricting the first-stage structural function, h.

PROPOSITION 2. *If $V = C(h(Z, \eta), Z)$ does not depend on Z given η for a $C(X, Z)$ that is smooth and almost surely invertible in X, then $\{h(\cdot, \eta)\}$ is a one-dimensional family of functions in Z.*

Remark. Identification of average structural functions or quantile structural functions for a given $X = x$ requires, in addition to conditional independence (4), that V has full support given $X = x$. In other words, the range of $C(X, Z)$ must be independent of X.

Remark. Assume that almost surely $h(Z, \eta_1) \neq h(Z, \eta_2)$ for independent draws Z, η_1, η_2 from the respective distributions of Z and η. Then the family of functions $\{h(., \eta)\}$ is one-dimensional if and only if it is possible to predict the counterfactual X under manipulation of Z from knowledge of X and Z. This possibility is a much stronger requirement than the possibility of identifying the ASF or QSF for the first stage relationship, which follows immediately from exogeneity of Z. The counterfactual outcome setting $Z = z_0$, $h(z_0, \eta)$ is used as a control function in the proof of Proposition 3 below.

Remark. If invertibility is dropped from the assumptions of Proposition 2, one-dimensionality of the family $\{h(\cdot, \eta)\}$ does not necessarily follow, but neither does

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

conditional independence (4). For example, if $C = const.$, then conditional independence (5) holds, but (4) does not necessarily hold.

The reverse of Proposition 2 is also true.

PROPOSITION 3. *If $\{h(.,\eta)\}$ is a one-dimensional family of functions in Z and almost surely $h(Z,\eta_1) \neq h(Z,\eta_2)$ for independent draws Z,η_1,η_2 from the respective distributions of Z and η, then there exists a control function $V = C(h(Z,\eta),Z)$ that does not depend on Z given η.*

Remark. If the family $\{h(.,\eta)\}$ is not only one-dimensional but also monotonic in unobserved heterogeneity, that is,

$$h(z_1,\eta_1) > h(z_1,\eta_2) \Leftrightarrow h(z_2,\eta_1) > h(z_2,\eta_2) \; \forall \; z_1,z_2,\eta_1,\eta_2, \tag{15}$$

then $C(X,Z) = F(X|Z)$ is the same control function as the one constructed in the proof of Proposition 3, in that there is an invertible mapping between the two. If monotonicity fails, however, $C(X,Z) = F(X|Z)$ cannot satisfy the sufficient condition of Proposition 1.

Remark. It follows from Proposition 2 that in the random coefficient example of Section 2, no control function satisfying the sufficient condition of Proposition 1 and invertibility in X can exist. The family of functions

$$h(Z,\eta_1,\eta_2) = \eta_1 + \eta_2 Z \tag{16}$$

assumed in the random coefficient model is two-dimensional, which implies that we cannot predict the counterfactual X under a manipulation setting $Z = z$, $h(z,\eta)$, for a given observational unit from X and Z alone.

3.2. Control Functions Satisfying Conditional Independence

Next we will consider the more general case of control functions satisfying conditional independence (4), which is required to identify $\mathrm{E}_{\epsilon|V}[g(x,\epsilon)|V]$ by $\mathrm{E}[Y|X=x,V]$.

PROPOSITION 4. *There exists a control function $V = C(X,Z)$ such that conditional independence (4) holds and that is invertible in Z if and only if $P(\epsilon|X,Z)$ is an at most one-dimensional family of distributions that is not constant in Z if it is not constant.*

Remark. If C is not invertible in Z, the following situation is theoretically possible: The family of conditional distributions $p(\epsilon|Z,X)$ is two-dimensional. The conditional support of (X,Z) given V, $\mathbb{XZ}(V)$, is comprised of a discrete set of points given X. Hence $p(\epsilon|X,V)$ is a mixture of $p(\epsilon|Z,X)$ over a discrete set of points Z. None of the components of this mixture is constant as X varies and Z

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

covaries to remain within the manifold $\mathbb{XZ}(V)$. Nevertheless, the changes in the components cancel exactly, implying that $p(\epsilon|X, V)$ is constant in X.

Intuitively, such canceling seems a highly nongeneric phenomenon and of little practical relevance. We do not have results, however, precluding this possibility in the absence of invertibility of C in Z.

Remark. The theorem only characterizes conditions for the existence of a control function. It does not give conditions for identifiability of C itself.

In the random coefficient model of Section 2, the necessary condition of Proposition 4 is not fulfilled in general. We have

$$\epsilon|X, Z \sim N\left(\mu_\epsilon + (X - \mu_{\eta_1} - \mu_{\eta_2} Z)\frac{\mathrm{Cov}(X, \epsilon|Z)}{\mathrm{Var}(X|Z)}, \mathrm{Var}(\epsilon) - \frac{\mathrm{Cov}^2(X, \epsilon|Z)}{\mathrm{Var}(X|Z)}\right), \tag{17}$$

which is a two-dimensional family as long as $\mathrm{Cov}(X, \epsilon|Z)$ is not identical 0 and $\frac{\mathrm{Cov}^2(X,\epsilon|Z)}{\mathrm{Var}(X|Z)}$ is not constant in Z, i.e., so long as

$$\frac{(\Sigma_{\eta_1,\epsilon} + Z\Sigma_{\eta_2,\epsilon})^2}{\Sigma_{\eta_1,\eta_1} + 2Z\Sigma_{\eta_1,\eta_2} + Z^2\Sigma_{\eta_2,\eta_2}}$$

depends on Z. Since this is the case for generic Σ, the following corollary holds.

COROLLARY 1. *There exists no control function invertible in Z in the generic random coefficient model of Section 2 such that conditional independence (4) holds.*

3.3. Higher-Dimensional X and Z

The results of the previous sections extend to the case of higher-dimensional X and Z. In particular, if we allow $X \in \mathbf{R}^k$ and $Z \in \mathbf{R}^l$ with $l \geq k$, the following generalization of Proposition 2 holds.

PROPOSITION 5. *If $V = C(h(Z, \eta), Z)$ does not depend on Z given η for a $C(X, Z)$ that is smooth and almost surely invertible in X, then $\{h(., \eta)\}$ is a k-dimensional family of functions in Z.*

The proof is analogous to the one-dimensional case. Similarly, for Proposition 3 we have the generalization below.

PROPOSITION 6. *If $\{h(., \eta)\}$ is a k-dimensional family of functions in Z and almost surely $h(Z, \eta_1) \neq h(Z, \eta_2)$ for independent draws Z, η_1, η_2 from the respective distributions of Z and η, then there exists a control function $V = C(h(Z, \eta), Z)$ that does not depend on Z given η.*

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

Finally, since none of the arguments leading to Proposition 4 depends on the dimensionality of X or Y, we get the generalization below.

PROPOSITION 7. *There exists a control function $V = C(X, Z)$ such that conditional independence (4) holds and that is invertible in Z if and only if $P(\epsilon|X, Z)$ is an at most l-dimensional family of distributions that is not constant in Z if it is not constant.*

4. CONCLUSION

This note characterizes triangular models for which control functions satisfying conditional independence and support requirements exist. These characterizations seem restrictive and will generally not be fulfilled. In particular, Proposition 2 states that having a control function that is a function of unobserved heterogeneity η requires a one-dimensional first stage family of structural functions.

Examples of such one-dimensional families include (i) families that are monotonic in unobserved heterogeneity, as in Imbens and Newey (2009); (ii) models with $X = h(|Z - \eta|)$, which could describe the loss from missing an unknown target η; and (iii) multiplicative families of the form $X = h(Z) \cdot \eta$, where h is of nonconstant sign. An economic example of (iii) is an income equation where X is income, $h(Z)$ is the (possibly negative) amount of some asset that an individual owns, and η is the rate of return. The characterizations proven in this paper show, however, that while such alternative families could be considered, no less restrictive family will allow construction of a control function. That is, there is no scope for generalization in conditions beyond Imbens and Newey (2009).

NOTES

1. Suppose the object of interest is $\widetilde{g}(x) = \text{argmin}_{\check{g}} \mathrm{E}_\epsilon[\rho(g(x, \epsilon), \check{g})]$ for a loss function $\rho(Y, \widetilde{g})$. The function $\widetilde{g}$ is identified under the same conditions as the ASF. To show this, replace Y with ρ in equation (6).
2. I thank Bryan Graham for this motivation.
3. If $\mu_{\eta_2} > 0$, then $Z \leq -\Sigma_{\eta_1,\eta_2} / \Sigma_{\eta_2,\eta_2}$ is sufficient, though not necessary.

REFERENCES

Blundell, R. & J. Powell (2003) Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, vol. 2, pp. 655–679.

Imbens, G. (2007) Nonadditive models with endogenous regressors. *Econometric Society World Congress Series* 75, 17–46.

Imbens, G. & W. Newey (2009) Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77, 1481–1512.

Newey, W., J. Powell, & F. Vella (1999) Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67, 565–603.

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press

APPENDIX: Proofs

Proof of Proposition 1. This is immediate from independence of Z and (η, ϵ). By assumption we can write V as a function of η, and therefore

$$Z|(V(\eta), \epsilon) \sim Z. \tag{A.1}$$

Hence, the conditional distribution of Z given $V(\eta), \epsilon$ does not depend on ϵ, implying conditional independence (5). ■

Proof of Proposition 2. Invertibility in X and smoothness of C imply that the range of $V = C(X, Z)$ is a one-dimensional, smooth manifold for a given Z. Since V is a function of η only, its range is independent of Z. Hence the range of V is a one-dimensional, smooth manifold. Invertibility and smoothness of C imply further that we can define a function $\tilde{h}$ such that $X = \tilde{h}(Z, V)$.

These assertions are true for many "reduced form" representations of the first stage such as regression residuals or conditional quantiles. However, the assumption that V does not depend on Z given η makes the first stage "structural" in the sense that we can write

$$h(Z, \eta) = \tilde{h}(Z, V(\eta)). \tag{A.2}$$

Because V is one-dimensional, this is a one-dimensional family of functions in Z. ■

Proof of Proposition 3. Since $\{h(., \eta)\}$ is a one-dimensional family of functions, we can assume without loss of generality that η has its support in $\mathbf{R}$. Pick a generic z_0 from the distribution of Z, and define $C(X, Z) = h(z_0, h^{-1}(Z, X))$, where the inverse is understood with respect to the η argument of h holding Z fixed. This inverse is well defined by the nonconstancy of h in η and the one-dimensionality of η. By definition, $C(h(Z, \eta), Z) = h(z_0, \eta)$, which is a function of η alone. ■

Proof of Proposition 4. Consider the family of conditional distributions of ϵ given Z, X. This is an at most two-dimensional family, indexed by a parameter that shall be denoted $\theta(Z, X)$, that is, $p(\epsilon|Z, X) =: p(\epsilon, \theta(Z, X))$. The distribution of ϵ given X, V is in general a mixture over Z of $p(., \theta(Z, X))$ for Z such that $C(X, Z) = V$. If C is invertible in Z, no mixing takes place, and this reduces to $p(., \theta(Z, X))$ for $Z = C^{-1}(X, V)$. In this case, conditional independence (4) is equivalent to constancy of $\theta(Z, X)$ on the manifold

$$\mathbb{XZ}(V) := \{(x, z) : C(x, z) = V, (x, z) \in supp(X, Z)\}. \tag{A.3}$$

Hence θ could be written as a function of C, which implies that the dimensionality of the range of θ is no higher than the dimensionality of the range of C. The range of C, however, cannot be of dimensionality larger than the dimension of Z if C has full range *given* X, as required for identification of the ASF and implied by the invertibility of C. This implies that the range of θ is at most one-dimensional. ■

https://doi.org/10.1017/S0266466610000460 Published online by Cambridge University Press