cemmap

centre for microdata methods and practice

Heterogenous coefficients, discrete instruments and identification of treatment effects

Whitney K. Newey
Sami Stouli

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP66/18

HETEROGENOUS COEFFICIENTS, DISCRETE INSTRUMENTS, AND IDENTIFICATION OF TREATMENT EFFECTS

WHITNEY K. NEWEY† AND SAMI STOULI§

Abstract. Multidimensional heterogeneity and endogeneity are important features of a wide class of econometric models. We consider heterogenous coefficients models where the outcome is a linear combination of known functions of treatment and heterogenous coefficients. We use control variables to obtain identification results for average treatment effects. With discrete instruments in a triangular model we find that average treatment effects cannot be identified when the number of support points is less than or equal to the number of coefficients. A sufficient condition for identification is that the second moment matrix of the treatment functions given the control is nonsingular with probability one. We relate this condition to identification of average treatment effects with multiple treatments.

Keywords: Heterogeneity, discrete instruments, control functions, nonseparable models, random coefficients.

1. Introduction

Nonseparable and/or multidimensional heterogeneity is important. It is present in discrete choice models as in McFadden (1973) and Hausman and Wise (1978). Multidimensional heterogeneity in demand functions allows price and income elasticities to vary over individuals in unrestricted ways, e.g., Hausman and Newey (2016) and Kitamura and Stoye (2016). It allows general variation in production technologies. Treatment effects that vary across individuals require intercept and slope heterogeneity.

Endogeneity is often a problem in these models because we are interested in the effect of an observed choice, or treatment variable on an outcome and the choice or treatment variable is correlated with heterogeneity. Control variables provide an important means of controlling for endogeneity with multidimensional heterogeneity. A control

Date: November 24, 2018.

† Department of Economics, MIT, wnewey@mit.edu.

§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

variable is an observed or estimable variable that makes heterogeneity and treatment independent when it is conditioned on. Observed covariates serve as control variables for treatment effects (Rosenbaum and Rubin, 1983). The conditional cumulative distribution function (CDF) of a choice variable given an instrument can serve as a control variable in triangular economic models (Imbens and Newey, 2009).

In fully nonparametric, nonseparable models identification of average or quantile treatment effects requires a full support condition, that the support of the control variable conditional on the treatment variable is equal to the marginal support of the control variable. This restriction is often not satisfied in practice; e.g., see Imbens and Newey (2009) for Engel curves. In triangular models the full support cannot hold when all instruments are discrete and the treatment variable is continuous.

One approach to this problem is to focus on identified sets for objects of interest, as for quantile effects in Imbens and Newey (2009). Another approach is to consider restrictions on the model that allow for point identification. Florens, Heckman, Meghir, and Vytlacil (2008) gave identification results when the outcome equation is a polynomial in the endogenous variable. Torgovitsky (2015) and D'Haultfœuille and Février (2015) gave identification results when there is only scalar heterogeneity in the outcome equation.

In this paper we give identification results when the outcome function is a linear combination of known functions of a treatment that are not necessarily polynomials. The coefficients in this linear combination are allowed to be heterogenous in unrestricted ways. We give identification results for average treatment effects in triangular models with discrete instruments. We find that a necessary condition for identification is that the number of support points of the discrete instruments is at least as large as the number of known functions of treatment in the outcome function. A sufficient identification condition is that the second moment matrix of the known functions conditional on the control function is nonsingular with probability one. We obtain these results from the implied varying coefficient structure of the regression of outcome on the treatment and control variables. We also use this approach to give identification results for average treatment effects with multidimensional treatments.

These results extend Florens, Heckman, Meghir, and Vytlacil (2008) in allowing for nonpolynomial functions of the treatment variable in the outcome equation and in allowing for discrete instruments. The results also show that it is possible to identify the average treatment effect when there is multidimensional heterogeneity and discrete instruments, in this way going beyond Torgovitsky (2015) and D'Haultfœuille and Février (2015). We also contribute to the literature on nonseparable models by giving identification results based on the conditional nonsingularity of the second moment matrix of functions of the treatment variable.

In Section 2 we introduce the model. In Section 3 we give the main identification result and the key implications of the identification condition. In Section 4 we discuss estimation of the model. Section 5 concludes. The proofs of all the results are given in the Appendix.

2. The Model

Let Y denote an outcome variable of interest, X an endogenous treatment, and ε a structural disturbance vector of finite dimension. We consider the heterogenous coefficients model

$$Y = p(X)'\varepsilon, \tag{2.1}$$

where $p(X)$ is a vector of known functions. This model is linear in the known functions $p(X)$ of the endogenous variables with coefficients ε that need not be independent of X. The coefficients ε characterise how $p(X)$ affects Y and can vary over individuals. This model generalises Florens, Heckman, Meghir, and Vytlacil (2008) to allow $p(X)$ to be any functions of X rather than just powers of X. When $p(X)$ is a vector of approximating functions such as splines or wavelets this model can be viewed as an approximation to a general nonseparable model $Y = g(X, \varepsilon)$ where ε are varying coefficients in an expansion of $g(x, \varepsilon)$ in $p(x)$, as in Hausman and Newey (2016). In this paper we take $p(X)'\varepsilon$ to be a correct model.

We consider the use of control variables to identify interesting objects associated with the function $p(X)'\varepsilon$. We assume that the vector ε is mean independent of the endogenous variable X conditional on an observable or estimable control variable denoted V.

Assumption 1. *For the model in (2.1), there exists a control variable V such that $E[\varepsilon \mid X, V] = E[\varepsilon \mid V]$.*

This conditional mean independence property and the form of the structural function $p(X)'\varepsilon$ in (2.1) together imply that X is known to affect the control regression function

(CRF) of Y given (X, V), $E[Y \mid X, V]$, only through the vector of known functions $p(X)$:

$$E[Y \mid X, V] = p(X)'E[\varepsilon|X, V] = p(X)'E[\varepsilon|V] = p(X)'q_0(V), \; q_0(V) \equiv E[\varepsilon|V]. \tag{2.2}$$

This control variable regression generalises that of Chernozhukov, Fernandez-Val, Newey, Stouli, and Vella (2017) to allow $q_0(V)$ to be a vector of unknown functions of V rather than a linear combination of finitely many known transformations of V. This restricted nonparametric regression is of the varying coefficients type considered by Cai, Das, Xiong, and Wu (2006).

An important kind of control variable arises in a triangular system where an instrumental variable Z is excluded from the outcome equation (2.1) and where X is a scalar with

$$X = h(Z, \eta), \tag{2.3}$$

with $h(z, \eta)$ strictly monotonic in η. If (ε, η) is jointly independent of Z, Assumption 1 is satisfied in the triangular system (2.1)-(2.3) with $V = F_{X|Z}(X \mid Z)$, the CDF of X conditional on Z (Imbens and Newey, 2009). Alternatively, $V = F_{X|Z}(X \mid Z)$ is a control variable in this model under the weaker conditions that η is independent from Z and that ε be mean independent of Z conditional on η.

Theorem 1. *For the triangular system (2.1)-(2.3), if η is independent from Z and $E[\varepsilon \mid \eta, Z] = E[\varepsilon \mid \eta]$ then $E[\varepsilon \mid X, V] = E[\varepsilon \mid V]$.*

Additional exogenous covariates Z_1 can be incorporated straightforwardly in the model through the known functional component of the CRF. With covariates Z_1, the CRF takes the form

$$E[Y \mid X, Z_1, V] = p(X, Z_1)'q_0(V),$$

where $p(X, Z_1)$ is a vector of known functions of (X, Z_1). The addition of exogenous covariates does not affect the identification analysis and it is straightforward to incorporate them, so we do not include them explicitly in the rest of this paper.

An important special case of this model is treatment effects where $p(X)$ is a vector that includes a constant and dummy variables for various kinds of treatments. For example the Rosenbaum and Rubin (1983) treatment effects model is included as a special case where $X \in \{0, 1\}$ is a treatment dummy variable that is equal to one if treatment occurs and equals zero without treatment and

$$p(X) = (1, X)'.$$

In this case $\varepsilon = (\varepsilon_1, \varepsilon_2)$ is two dimensional with ε_1 giving the outcome without treatment and ε_2 being the treatment effect. Here the control variables in V would be observable variables with Assumption 1 holding, i.e., the coefficients $(\varepsilon_1, \varepsilon_2)$ are mean independent of treatment conditional on controls, which is the unconfoundedness assumption of Rosenbaum and Rubin (1983). Multiple treatments could also be allowed for by letting X be a vector of dummy variables with each variable representing a different kind of treatment, e.g., Imbens (2000).

A central object of interest in model (2.1) is the average structural function (ASF) given by $\mu(x) \equiv p(x)'E[\varepsilon]$; see Blundell and Powell (2003) and Wooldridge (2005). When $X \in \{0, 1\}$ is a dummy variable for treatment $\mu(0)$ gives the average outcome for those not treated and $\mu(1)$ the average outcome for those who are treated, with $\mu(1) - \mu(0)$ being the average treatment effect. When X is continuously distributed $\partial\mu(x)/\partial x = [\partial p(x)/\partial x]'E[\varepsilon]$ gives a derivative version of the average treatment effect.

The ASF can be expressed as a known linear combination of $E[q_0(V)]$ from equation (2.2). By iterated expectations

$$p(x)'E[q_0(V)] = p(x)'E[E[\varepsilon \mid V]] = \mu(x).$$

We note that identification of the ASF requires integrating over the marginal distribution of the control variable V. There are other interesting structural objects that do not rely only on the marginal distribution of V. For example, the average derivative of the structural function is

$$\begin{aligned} E\left[\frac{\partial\{p(X)'\varepsilon\}}{\partial x}\right] &= E\left[E\left[\frac{\partial\{p(X)'\varepsilon\}}{\partial x} \mid X, V\right]\right] = E\left[\left\{\frac{\partial p(X)}{\partial x}\right\}' E[\varepsilon \mid X, V]\right] \\ &= E\left[\left\{\frac{\partial p(X)}{\partial x}\right\}' E[\varepsilon \mid V]\right] = E\left[\frac{\partial E[Y \mid X, V]}{\partial x}\right], \end{aligned}$$

as shown in Imbens and Newey (2009) for general nonseparable models. This object and others like it, including the local average response of Altonji and Matzkin (2005), do not require the full support condition for identification in general nonseparable models. For this reason we focus our identification results on the ASF where we use the heterogenous coefficients structure to weaken the full support condition for identification.

3. Identification

3.1. **Main Results.** One main contribution of this paper is to highlight and show that in the heterogenous coefficients model we consider the ASF is identified under the following condition:

Assumption 2. *$E\left[\|\varepsilon\|^4\right] < \infty$, $E\left[\|p(X)\|^4\right] < \infty$, and $E\left[p(X)\,p(X)' \mid V\right]$ is nonsingular with probability one.*

This condition is sufficient for identification of the unknown function $q_0(V)$.

Theorem 2. *If Assumptions 1-2 hold then $q_0(V)$ is identified.*

We discuss below conditions for nonsingularity of $E\left[p(X)\,p(X)' \mid V\right]$. All those conditions are sufficient for identification of $q_0(V)$, including those that allow for discrete valued instrumental variables in triangular systems. We also note that identification of $q_0(V)$ means uniqueness on a set of V having probability one. Thus the ASF will be identified as

$$\mu(X) = p(X)'\,E[q_0(V)], \tag{3.1}$$

with probability one. In other words, the ASF is identified because $p(X)$ is a known function and $q_0(V)$ is identified, and hence $E[q_0(V)]$ also is.

Theorem 3. *If Assumptions 1-2 hold then the ASF is identified.*

The identification analysis applies to other control regressions and model specifications. First, Theorems 2 and 3 also apply to models with CRF of the form $E[Y \mid X, V] = p_0(X)'q(V)$, where $p_0(X)$ is now unknown and V affects $E[Y \mid X, V]$ only through a vector of known functions $q(V)$. Identification of $p_0(X)$ then requires that $E\left[q(V)\,q(V)' \mid X\right]$ be nonsingular with probability one. Second, Theorem 2 directly applies to other control regressions, such as the control conditional quantile function $Q_{Y|XV}(u \mid X, V) = p(X)'q_u(V)$, $u \in (0,1)$, and the control CDF $F_{Y|XV}(y \mid X, V) = p(X)'q_y(V)$. Third, the identification condition also implies identification of control regressions where the functional form of both how X and V affect the outcome is restricted, under the weaker condition that it holds on a set of V having positive probability. A detailed exposition of all these alternative control regression specifications and the corresponding average, quantile and distribution structural functions is given in Newey and Stouli (2018).

3.2. **Discussion of the Identification Condition.** The nonsingularity condition on $E[p(X)p(X)' \mid V]$ allows for the support of X conditional on V to be discrete. Under this condition, the heterogenous coefficients form of model (2.1) and identification of $q_0(V)$ together imply uniqueness of the CRF on a set of (X, V) having probability one. Therefore the integral in the definition (3.1) of the ASF is well-defined because integration then occurs over a range of v values conditional on X where the CRF is identified. Stronger sufficient conditions for identification are full support (Imbens and Newey, 2009), that the support of V conditional on X equals the marginal support of V, and measurable separability (Florens, Heckman, Meghir, and Vytlacil, 2008), that any function of X equal to a function of V with probability one must be equal to a constant with probability one. Both conditions require X to have continuous support conditional on V. The formulation of an identification condition in terms of the conditional second moment matrix of $p(X)$ thus represents a substantial weakening of the conditions previously available in the literature.

When $X \in \{0, 1\}$ and $p(X) = (1, X)'$, the identification condition becomes the standard condition for the treatment effect model

$$Y = \varepsilon_1 + \varepsilon_2 X, \quad E[\varepsilon \mid X, V] = E[\varepsilon \mid V], \quad \varepsilon \equiv (\varepsilon_1, \varepsilon_2)'.$$

The identification condition is that the conditional variance matrix of $(1, X)'$ given V is nonsingular with probability one, which is the same as

$$\text{(3.2)} \qquad \mathrm{Var}(X \mid V) = P(V)[1 - P(V)] > 0, \quad P(V) \equiv \Pr(X = 1 \mid V),$$

with probability one, where $P(V)$ is the propensity score. Here we can see that the identification condition is the same as $0 < P(V) < 1$ with probability one, which is the standard identification condition.

With multiple treatments letting X denote a vector of dummy variables $X(t)$, $t \in \mathcal{T} \equiv \{1, \ldots, T\}$, taking value one if treatment t occurs and zero otherwise, the identification condition is that the conditional second moment matrix of

$$p(X) = (1, X(1), \ldots, X(T))'$$

given V is nonsingular with probability one. For the case where V is observable, Graham and Pinto (2018) address similar issues in independent work. Considerable simplification occurs when the treatments are mutually exclusive, allowing for the characterisation of a necessary and sufficient condition for nonsingularity of $E[p(X)p(X)' \mid V]$ that generalises (3.2).

Theorem 4. *Suppose that* $\Pr(X(t) = 1 \mid V) > 0$ *for each* $t \in \mathcal{T}$. *With mutually exclusive treatments,* $E[p(X)p(X)' \mid V]$ *is nonsingular with probability one if and only if*

$$1 - \Sigma_{s=1}^{T} \Pr(X(s) = 1 \mid V) > 0$$

with probability one.

In triangular systems with control variable $V = F_{X|Z}(X \mid Z)$, our identification condition can equivalently be stated in terms of the first stage representation $X = Q_{X|Z}(V \mid Z)$ and the instrument Z. By independence of V from Z, a sufficient condition for identification is that

$$E[p(Q_{X|Z}(v \mid Z))p(Q_{X|Z}(v \mid Z))']$$

be nonsingular for almost every $v \in \mathcal{V}$. When Z is discrete with support $\mathcal{Z} = \{z : \Pr(Z = z) \geq \delta > 0\}$ of finite cardinality $|\mathcal{Z}|$, a necessary condition for nonsingularity is then that the set $\mathcal{Q}(V)$ of distinct values[1] of $z \mapsto Q_{X|Z}(V \mid z)$ has cardinality $|\mathcal{Q}(V)|$ greater than or equal to $J \equiv \dim(p(X))$ with probability one.

Theorem 5. *Suppose that* $|\mathcal{Z}| < \infty$. *Then* $E[p(X)p(X)' \mid V]$ *is nonsingular with probability one only if* $\Pr[|\mathcal{Q}(V)| \geq J] = 1$.

Theorem 5 formalises the intuitive notion that the complexity of the model as measured by the dimension of its known functional component $p(X)$ is restricted by the cardinality of the set of instrumental values. Thus only when $p(X) = (1, X)'$ can identification be achieved in the presence of a binary instrument. A more primitive condition for identification in this case is that a change in the value of the instrument shifts the value of the conditional quantile function $z \mapsto Q_{X|Z}(V \mid z)$ with probability one.

Theorem 6. *Let* $p(X) = (1, X)'$. *If* $|\mathcal{Z}| = 2$ *and* $\Pr[Q_{X|Z}(V \mid z_1) \neq Q_{X|Z}(V \mid z_2)] = 1$, *then* $E\left[p(X)\, p(X)' \mid V\right]$ *is nonsingular with probability one.*

[1]Formally, for $v \in (0, 1)$, we define $\mathcal{Q}(v) = \left\{Q_{X|Z}(v \mid z_m)\right\}_{m \in \mathcal{M}(v)}$, where

$$\mathcal{M}(v) = \left\{m \in \{1, \ldots, |\mathcal{Z}|\} : Q_{X|Z}(v \mid z_m) \neq Q_{X|Z}(v \mid z_{m'}) \text{ for all } m' \in \{1, \ldots, |\mathcal{Z}|\}/\{m\}\right\}.$$

4. Estimation

The results of the previous Section lead to direct estimation methods for the heterogenous coefficients model we consider. One approach to making estimation feasible is through the approximation of the nonparametric component $q_0(V)$ by approximating functions such as splines or wavelets.

For the CRF specification $E[Y \mid X, V] = p(X)'q_0(V)$, we approximate each component $q_{0j}(V)$, $j \in \mathcal{J} \equiv \{1, \ldots, J\}$, of the unknown functional coefficient vector $q_0(V)$ by a linear combination of K basis functions $\psi^K = (\psi_1^K, \ldots, \psi_K^K)'$,

$$q_{0j}(V) \approx \sum_{k=1}^{K} b_{jk}\psi_k^K(V) = b_j'\psi^K(V), \quad j \in \mathcal{J}, \tag{4.1}$$

where $b_j = (b_{j1}, \ldots, b_{jK})'$, $j \in \mathcal{J}$, which yields an approximation of the form

$$E[Y \mid X, V] = p(X)'q_0(V) \approx \sum_{j=1}^{J} \left\{b_j'\psi^K(V)\right\} p_j(X) = b'[p(X) \otimes \psi^K(V)],$$

where $b = (b_1', \ldots, b_J')'$. Such an approximation is well-defined under our conditions with $b = b_{\mathrm{LS}}^K$, the coefficient vector of a least squares regression of Y on $p(X) \otimes \psi^K(V)$,

$$b_{\mathrm{LS}}^K \equiv \arg\min_{b \in \mathbb{R}^{JK}} E\left[\left\{Y - b'[p(X) \otimes \psi^K(V)]\right\}^2\right]. \tag{4.2}$$

The proposed approximation is valid for the CRF $E[Y \mid X, V]$ if the specified basis functions satisfy the following condition.

Assumption 3. *For all K, $E\left[||\psi^K(V)||^2\right] < \infty$, $E[\psi^K(V)\psi^K(V)']$ exists and is nonsingular, and, for any J vector of functions $a(V)$ with $E[||a(V)||^2] < \infty$, there are $K \times 1$ vectors φ_j^K, $j \in \mathcal{J}$, such that as $K \to \infty$, $E[\sum_{j=1}^J \{a_j(V) - \psi^K(V)'\varphi_j^K\}^2] \to 0$.*

Under this assumption we have that, as $K \to \infty$,

$$E\left[\left\{E[Y \mid X, V] - [p(X) \otimes \psi^K(V)]'b_{\mathrm{LS}}^K\right\}^2\right] \to 0. \tag{4.3}$$

Therefore $E[Y \mid X, V]$ can be approximated arbitrarily well by increasing the number of terms in the approximate specification (4.1)

Theorem 7. *Suppose that $E\left[\|q_0(V)\|^2\right] < \infty$ and $\sup_{v \in \mathcal{V}} E[||p(X)||^2 \mid V = v] \leq C$ for some finite constant C. Then (4.3) holds under Assumptions 1-3.*

An estimator for the CRF is given by taking the sample analog in (4.2), upon substituting for the control variable V by its estimated version when it is unobservable. The properties of the corresponding ASF estimator, including convergence rates and asymptotic normality, have been extensively analysed by Imbens and Newey (2002) for the general case where $p(X)$ and $q(V)$ both are increasing sequences of splines or power series approximating functions and the vector of regressors is of the kronecker product form we consider (cf. Theorems 6–8 in Imbens and Newey, 2002). Their analysis accounts for a first step nonparametric estimate of the control variable, and their results directly apply to the simpler case we consider here where the dimension of $p(X)$ is fixed, including when V is observable. In particular, we find that the convergence rate for the ASF in the model is solely determined by the rate of the first step estimator for the control variable. An immediate and remarkable corollary of this result is that in the model average treatment effects are estimable at a parametric rate when V is itself estimable at a parametric rate[2] or observable.

5. Conclusion

This paper introduces a new, transparent nonsingularity condition for the identification of models with heterogenous coefficients and endogenous treatments. We use this condition to give identification results that allow for discrete instruments in triangular systems with multidimensional unobserved heterogeneity. The approach applies to various types of treatment effects and model specifications, including average treatment effects with multiple treatments, and the model can be conveniently estimated by a series-based least squares estimator with well-understood properties.

Appendix A. Proof of Main Results

A.1. **Proof of Theorem 1.**

Proof. As in the proof of Theorem 1 in Imbens and Newey (2009), V is a one-to-one function of η. Then by equation (2.3), iterated expectations, and conditional mean

[2]Models that allow for estimation of the CDF $F_{X|Z}(X \mid Z)$ at a parametric rate can be formulated using quantile and distribution regression (Chernozhukov, Fernandez-Val, Newey, Stouli, and Vella, 2017) or dual regression (Spady and Stouli, 2018).

independence,

$$\begin{aligned}E[\varepsilon \mid X, V] = E[\varepsilon \mid h(Z,\eta),\eta] &= E[E[\varepsilon \mid \eta, Z] \mid h(Z,\eta),\eta] \\ &= E[E[\varepsilon \mid \eta] \mid h(Z,\eta),\eta] = E[\varepsilon \mid \eta] = E[\varepsilon \mid V],\end{aligned}$$

as claimed. □

A.2. **Proof of Theorem 2.**

Proof. Let $\lambda_{\min}(V)$ denote the smallest eigenvalue of $E[p(X)p(X)' \mid V]$. Suppose that $\bar{q}(V) \neq q_0(V)$ with positive probability on a set $\widetilde{\mathcal{V}}$, and note that $\lambda_{\min}(V) > 0$ on $\mathcal{V}$ by Assumption 2. Then

$$\begin{aligned}E\left[\left\{p(X)'\{\bar{q}(V) - q_0(V)\}\right\}^2\right] &= E\left[\{\bar{q}(V) - q_0(V)\}' E\left[p(X)p(X)' \mid V\right]\{\bar{q}(V) - q_0(V)\}\right] \\ &\geq E\left[\|\bar{q}(V) - q_0(V)\|^2 \lambda_{\min}(V)\right] \\ &\geq E\left[1(V \in \mathcal{V} \cap \widetilde{\mathcal{V}})\,\|\bar{q}(V) - q_0(V)\|^2 \lambda_{\min}(V)\right]\end{aligned}$$

By definition $\Pr(\widetilde{\mathcal{V}}) > 0$ and $\widetilde{\mathcal{V}} \subseteq \mathcal{V}$ so that $\widetilde{\mathcal{V}} \cap \mathcal{V} = \widetilde{\mathcal{V}}$. Thus the fact that $\|\bar{q}(V) - q_0(V)\|^2 \lambda_{\min}(V)$ is positive on $\widetilde{\mathcal{V}} \cap \mathcal{V}$ implies

$$E\left[1(V \in \mathcal{V} \cap \widetilde{\mathcal{V}})\,\|\bar{q}(V) - q_0(V)\|^2 \lambda_{\min}(V)\right] > 0.$$

We have shown that, for $\bar{q}(V) \neq q_0(V)$ with positive probability on a set $\widetilde{\mathcal{V}}$,

$$E\left[\left\{p(X)'\{\bar{q}(V) - q_0(V)\}\right\}^2\right] > 0,$$

which implies $p(X)'\bar{q}(V) \neq p(X)'q_0(V)$. Therefore, $q_0(V)$ is identified from $E[Y \mid X, V]$. □

A.3. **Proof of Theorem 3.**

Proof. Under Assumption 2, $q_0(V)$ is identified by Theorem 2. The result then follows from the argument in the text. □

A.4. **Proof of Theorem 4.**

Proof. For each $t \in \mathcal{T}$ and a vector $w \in \mathbb{R}^t$, let $\operatorname{diag}(w)$ denote the $t \times t$ diagonal matrix with diagonal elements $w_1, \ldots, w_t$, and define $X_t = (X(1), \ldots, X(t))'$ and

$p_t(X) = (1, X_t')'$. For mutually exclusive treatments $E[p_t(X)p_t(X)' \mid V]$ is of the form

$$E[p_t(X)p_t(X)' \mid V] = \begin{bmatrix} 1 & E[X_t' \mid V] \\ E[X_t \mid V] & \operatorname{diag}(E[X_t \mid V]) \end{bmatrix}. \tag{A.1}$$

For each $t \in \mathcal{T}$, using that $E[X_t \mid V] = \Pr(X(t) = 1 \mid V) > 0$, we have that $E[p_t(X)p_t(X)' \mid V]$ is positive definite if and only if the Schur complement of $\operatorname{diag}(E[X_t \mid V])$ in (A.1) is positive definite (Boyd and Vandenberghe, 2004, Appendix A.5.5.), i.e., if and only if

$$1 - E[X_t' \mid V] \operatorname{diag}(E[X_t \mid V])^{-1} E[X_t \mid V] = 1 - \Sigma_{s=1}^{t} E[X(s) \mid V] > 0.$$

The result now follows by $1 - \Sigma_{s=1}^{t} E[X(s) \mid V] \geq 1 - \Sigma_{s=1}^{T} \Pr(X(s) = 1 \mid V)$ and from the fact that a matrix is positive definite if and only if all its principal minors have strictly positive determinant. $\square$

A.5. **Proof of Theorem 5.**

Proof. By definition of $\mathcal{Z}$ we have that $\Pr(Z = z_m) \geq \delta > 0$ for $m \in \{1, \dots, |\mathcal{Z}|\}$. Thus, upon using the identity $X = Q_{X|Z}(V \mid Z)$ and by independence of V from Z, for $v \in (0, 1)$,

$$E[p(X)p(X)' \mid V = v] = \sum_{m=1}^{|\mathcal{Z}|} \left\{ p(Q_{X|Z}(v \mid z_m)) p(Q_{X|Z}(v \mid z_m))' \right\} \times \Pr(Z = z_m),$$

is a sum of $|\mathcal{Q}(v)| \leq |\mathcal{Z}|$ rank one $J \times J$ distinct matrices which is singular if $|\mathcal{Q}(v)| < J$. Thus if $|\mathcal{Q}(V)| < J$ with positive probability, then $E[p(X)p(X)' \mid V]$ is singular with positive probability. Therefore $E[p(X)p(X)' \mid V]$ is nonsingular with probability one only if $\Pr[|\mathcal{Q}(V)| \geq J] = 1$. $\square$

A.6. **Proof of Theorem 6.**

Proof. By assumption $Q_{X|Z}(V \mid Z)$ takes two values $Q_X(V \mid z_1)$ and $Q_X(V \mid z_2)$, $Q_{X|Z}(V \mid z_1) \neq Q_{X|Z}(V \mid z_2)$, with probability one. Moreover, by definition of $\mathcal{Z}$ and by independence of V from Z, we have that $\Pr(Z = z_m) = \Pr(Z = z_m \mid V) \geq \delta > 0$, $m = 1, 2$, with probability one. It follows that $\operatorname{Var}(Q_{X|Z}(V \mid Z) \mid V) > 0$ with probability one. Therefore by $p(X) = (1, X)'$, we have that $\det\left(E\left[p(X)\, p(X)' \mid V\right]\right) = \operatorname{Var}(Q_{X|Z}(V \mid Z) \mid V) > 0$ with probability one. $\square$

A.7. **Proof of Theorem 7.**

Proof. Let $q^K(V,b) = (q_1^K(V,b_1),\ldots,q_J^K(V,b_J))'$, where $q_j^K(V,b_j) \equiv \sum_{k=1}^K b_{jk}\psi_k^K(V)$, $j \in \mathcal{J}$. Model (2.1) and Assumptions 1-2 together imply that $E[Y \mid X,V] = p(X)'q_0(V)$, where $q_0(V)$ is unique with probability one by Theorem 2. Thus, for all $b \in \mathbb{R}^{JK}$,

$$
\begin{aligned}
E\left[\left\{E[Y \mid X,V] - b'[p(X)\otimes\psi^K(V)]\right\}^2\right] &= E\left[\left\{p(X)'q_0(V) - b'[p(X)\otimes\psi^K(V)]\right\}^2\right] \\
&= E\left[\left\{p(X)'\left[q_0(V) - q^K(V,b)\right]\right\}^2\right]. \qquad \text{(A.2)}
\end{aligned}
$$

Using that b_{LS}^K in (4.2) also satisfies

$$
b_{\mathrm{LS}}^K = \arg\min_{b\in\mathbb{R}^{JK}} E\left[\left\{E[Y \mid X,V] - b'[p(X)\otimes\psi^K(V)]\right\}^2\right],
$$

equation (A.2) implies that

$$
b_{\mathrm{LS}}^K = \arg\min_{b\in\mathbb{R}^{JK}} E\left[\left\{p(X)'\left[q_0(V) - q^K(V,b)\right]\right\}^2\right]. \qquad \text{(A.3)}
$$

Thus if, as $K\to\infty$,

$$
E\left[\left\{p(X)'\left[q_0(V) - q^K(V,b_{\mathrm{LS}}^K)\right]\right\}^2\right] \to 0,
$$

then the result follows.

Define

$$
\widetilde{b}^K \equiv \arg\min_{b\in\mathbb{R}^{JK}} E\left[||q_0(V) - q^K(V,b)||^2\right].
$$

We have that, as $K\to\infty$,

$$
\begin{aligned}
0 \le E\left[\left\{p(X)'\left[q_0(V) - q^K(V,\widetilde{b}^K)\right]\right\}^2\right] &\le E\left[||p(X)||^2\,||q_0(V) - q^K(V,\widetilde{b}^K)||^2\right] \\
&= E\left[E\left[||p(X)||^2 \mid V\right]\,||q_0(V) - q^K(V,\widetilde{b}^K)||^2\right] \\
&\le CE\left[||q_0(V) - q^K(V,\widetilde{b}^K)||^2\right] \to 0,
\end{aligned}
$$

by Cauchy-Schwarz, iterated expectations, uniform boundedness of $E[||p(X)||^2 \mid V = v]$ over $v \in \mathcal{V}$ and Assumption 3. Thus $\widetilde{b}^K$ is a minimiser of (A.3) for K large enough. We have that $E[p(X)p(X)' \mid V]$ is nonsingular with probability one by Assumption 2, and that $E[\psi^K(V)\psi^K(V)']$ is nonsingular by Assumption 3. Thus the matrix $E[\{p(X)\otimes\psi^K(V)\}\{p(X)\otimes\psi^K(V)\}']$ is nonsingular for each K by Theorem 3 in Newey and Stouli (2018). Therefore, $\widetilde{b}^K$ is the unique minimiser of (A.3) for K large enough. Conclude that $b_{\mathrm{LS}}^K = \widetilde{b}^K$ for K large enough, and the result follows. □

References

[1] Altonji, J. and Matzkin, R. (2005). Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica.* 73(4, July), pp.1053-1102.

[2] Blundell, R., and Powell, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric society monographs.* 36, pp. 312-357.

[3] Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press.

[4] Chernozhukov, V., Fernandez-Val, I. Newey, W., Stouli, S. and Vella, F. (2017). Semiparametric estimation of structural functions in non-separable triangular models. *eprint arXiv:1711.02184.*

[5] Cai, Z., Das, M., Xiong, H. and Wu, X. (2006). Functional coefficient instrumental variables models. *Journal of Econometrics.* 133, pp. 207-241.

[6] D'Haultfœuille, X. and Février, P. (2015). Identification of nonseparable triangular models with discrete instruments. *Econometrica* 83(3, May), pp. 1199-1210.

[7] 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(5, September), pp. 1191-1206.

[8] Graham, B. S. and Pinto, C. C. D. X. (2018) Semiparametrically efficient estimation of the average linear regression function. *eprint arXiv:1810.12511.*

[9] Hausman, J. A. and Newey, W. K. (2016) Individual heterogeneity and average welfare. *Econometrica.* 84(3, May), pp.1225-1248.

[10] Hausman, J. A. and Wise, D. (1978). A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica.* 46, pp. 403-426.

[11] Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika.* 87, pp. 706-710.

[12] Imbens, G. W. and Newey, W. K. (2002). Identification and estimation of triangular simultaneous equations models without additivity. NBER Technical Working Papers 0285.

[13] Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*. 77(5, September), pp. 1481-1512.

[14] Kitamura, Y. and Stoye, J. (2016). Nonparametric analysis of random utility models. *eprint arXiv:1606.04819*.

[15] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior, in: P. Zarambka (Ed.), Frontiers in econometrics. New York: Academic Press.

[16] Newey, W. and Stouli, S. (2018). Control variables, discrete instruments, and identification of structural functions. *eprint arXiv:1809.05706*

[17] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*. 70, pp.41-55.

[18] Spady, R. H. and Stouli, S. (2018). Dual regression. *Biometrika*. 105, pp. 1-18.

[19] Torgovitsky, A. (2015). Identification of nonseparable models using instruments with small support. *Econometrica*. 83(3, May), pp. 1185-1197.

[20] Wooldridge, J. M. (2005). *Identification and inference for econometric models*. Chapter Unobserved heterogeneity and the estimation of average partial effects, (pp. 27-55). Number 3. Cambridge University Press: Cambridge.