Endogeneity in Additive Nonparametric Models

Thomas Wiemann
University of Chicago

TA Discussion # 2

October 13, 2021

Introduction

Analysis of data with endogenous regressors is arguably the main contribution of econometrics to statistical science (according to Blundell and Powell, 2006).

Some key developments in the literature include the invention of TSLS by Theil (1953) (who is seemingly ridiculously under-cited), nonlinear parametric models with additive non-Gaussian errors (e.g., Amemiya, 1974; Hansen, 1982), and semi- and nonparametric models (e.g., Newey and Powell, 2003; Newey et al., 1999).

Today is meant as an introduction to identification in nonparametric models with endogeneity, as well as to some of the key difficulties associated with estimation.

The focus will be exclusively on *additive* nonparametric models. Hopefully, we can turn to non-additive nonparametric settings with endogeneity in future weeks.

An Additive Nonparametric Model

Consider the setting

$$Y = g_0(X) + U, \qquad E[U|Z] = 0, \tag{1}$$

where g_0 denotes the true, unknown structural function of interest, X is a vector of explanatory variables, and Z is a vector of instruments, and U is a disturbance.

Taking conditional expectations, we obtain

$$\pi(Z) := E[Y|Z] = E[g_0(X)|Z] = \int g_0(x)dF(x|Z), \tag{2}$$

where F denotes the conditional c.d.f of X given Z. (Under some technical conditions) π and F are unique functions of the data and are thus identified.

Two key questions:

- When is g_0 identified?
- When can we construct a consistent estimator for g_0?

Finite-Support Setting

Suppose that X and Z have joint distribution that is discrete with finite support. Take $\{\xi_j, j = 1, \ldots, J\}$ and $\{\vartheta_k, j = 1, \ldots, L\}$ denote the support points for X and Z, respectively, where

$$\begin{aligned} \pi &:= \text{vec}(E[y|Z = \vartheta_k]), \\ \mathbf{P} &:= [P(X = \xi_j|Z = \varphi_k)]. \end{aligned} \tag{3}$$

Note that π and $\mathbf{P}$ are identified as unique functions of the data.

Identifiablity of $\mathbf{g} := \text{vec}(g(\xi_j))$ is equivalent to uniqueness of a solution to

$$\pi = \mathbf{P}\mathbf{g}. \tag{4}$$

Hence, $\mathbf{g}$ is identified if and only if rank $P = J = \dim \mathbf{g}$. Notice that this requires $L = \dim \pi \geq \dim \mathbf{g} = J$.

When $\mathbf{g}$ is identified, it can be consistently estimated by replacing π and $\mathbf{P}$ by their sample analogues and solving for $\hat{\mathbf{g}}$ (when $J = L$) or using a minimum chi-square procedure (when $J < L$). See also Das (2005).

The Ill-Posed Inverse Problem

Bad news: The simple structure of the finite-support example does not generalize to X, Z being (partially) continuous.

Consider again the integral equation

$$\pi(Z) := E[Y|Z] = \int g_0(x) dF(x|Z). \tag{5}$$

This is a generalized version of a Fredholm equation of the first kind. The inverse problem for g_0 is well-posed if it has a unique solution g_0 for all π and this solution depends continuously on π. In this case, g_0 is stable w.r.t. small changes in π. If it is not well-posed, it is ill-posed.

It is this instability that is a key concern. Notice that as a consequence of ill-posedness, $E_n[Y|Z] \xrightarrow{p} E[Y|Z]$ and $F_n(X|Z) \xrightarrow{p} F(X|Z)$ *do not* imply $\hat{g}_n \xrightarrow{p} g_0$ (even if g_0 is identified!).

Intuition from Blundell and Powell (2006): This is a functional analogue to the problem of multicollinearity in linear regression, where large differences in regression coefficients can correspond to small differences in fitted values.

Nonparametric First Stage

So how can well-posedness (= identified + stable) be ensured?

Can restrict allowable functions g. E.g., $g_0(x) = x^\top \beta_0$. Then the integral equation reduces to

$$E[Y|Z] = E[X|Z]^\top \beta. \tag{6}$$

Notice

$$\begin{aligned} & E[Y|Z] = E[X|Z]^\top \beta \\ \Rightarrow \quad & E[X|Z]E[Y|Z] = E[X|Z]E[X|Z]^\top \beta \\ \Rightarrow \quad & E\left[E[X|Z]E[Y|Z]\right] = E\left[E[X|Z]E[X|Z]^\top\right]\beta, \end{aligned} \tag{7}$$

where $E\left[E[X|Z]E[X|Z]^\top\right]$ is invertible if $E[X|Z]$ is not perfectly collinear. Notice that stability is implied as long as the singular values of $E\left[E[X|Z]E[X|Z]^\top\right]$ are sufficiently far away from zero.

While this is semi-parametric, it hardly seems in the spirit of what we set out to do.

Completeness

For identification, Newey and Powell (2003) leave g_0 unrestricted and instead place restrictions on $F(X|Z)$. They leverage the notion of completeness in Z of the conditional distribution of X given Z, which is equivalent to identification of g_0:

Proposition 2.1 in Newey and Powell (2003): If equation (2) is satisfied, then g_0 is identified if and only if $\forall \delta(X) : E|\delta(X)| < \infty$, it holds that $E[\delta(X)|Z] = 0 \Leftrightarrow \delta(X) = 0$.

Loosely:

$$\begin{aligned} & \nexists \tilde{g} \neq g : \pi(Z) = \int \tilde{g}(x) dF(x|Z) \\ \Leftrightarrow \quad & \nexists \tilde{g} \neq g : \int (\tilde{g}(x) - g_0(x))\, dF(x|Z) = 0 \qquad (8) \\ \Leftrightarrow \quad & \nexists \delta \neq 0 : \int \delta(x) dF(x|Z) = 0 \end{aligned}$$

The questions then becomes: What kind of distributions $F(X|Z)$ are complete in Z *and* do they appear reasonable in economic applications?

Definition of completeness here.

Completeness (Contd.)

Newey and Powell (2003) give a sufficient condition for identification:

Theorem 2.2 in Newey and Powell (2003): If (2) is satisfied, with probability one conditional on Z the distribution of X is absolutely continuous with density

$$f(x|z) = s(x)t(z)\exp\left[\mu(z)^\top \tau(x)\right], \tag{9}$$

$s(x) > 0$, $\tau(x)$ is one-to-one in x, and the support of $\mu(z)$ is an open set, then g_0 is identified.

Here, the conditional distribution of X given Z is assumed to be an exponential family distribution (e.g., Gaussian).

But not all "reasonable" distributions are exponential family distributions. Gaussian mixtures, for example, are not in the exponential family.

Completeness (Contd.)

Of course, sufficient conditions are not (necessarily) necessary. So there might be hope.

Note that completeness is a restriction on the conditional distribution $F(X|Z)$, which is identified from the data. Canay et al. (2013) consider whether it is possible to test the H_0 that a completeness condition does not hold against the H_1 that it holds. They derive an impossibility results (under "commonly imposed" restrictions).

Importantly, nontestability of completeness does not imply that the assumption is false, but it does suggest that it is prudent to justify their use with alternative arguments in favor of their validity. E.g., does completeness hold for a sufficiently general class of distributions that are reasonable in economic applications?

Andrews (2017) gives examples of distributions that are L^2-complete.

Note also that completeness (or point identification) are not necessary for conducting nonparametric inference in models with endogeneity.

Estimation in Newey et al. (1999)

Suppose now that we can assume completeness of $F(X|Z)$ in Z such that g_0 is identified. We still need to worry about ill-posedness so that a consistent estimate can be obtained.

Newey and Powell (2003) assume g_0 belongs to a compact set. This approach eliminates the ill-posed inverse problem because it implies continuity of the inverse. (Arguments from functional analysis are necessary here. Very loosely: the integral is itself a map between two function spaces. If the domain (the space of g_0) is compact and the range (the space of π) is a metric space, then the inverse of the integral is continuous by, e.g., Theorem 26.6 in Munkres, 2000.)

The estimator is a bit involved (and ex ante appears sensitive to regularization parameters). We'll skip it here. See Section 3 in Newey and Powell (2003) for details.

There's a "modern" estimator in framework of Newey and Powell (2003) using deep neural networks: Hartford et al. (2017). There's no discussion of consistency or inference (as seems standard in ICML publications), but it's a fun read nevertheless.

Nonparametric control function

Consider the triangular system of equations of Newey et al. (1999):

$$\begin{aligned} Y &= g_0(X) + U \\ X &= \Pi_0(Z) + V \\ E[U|V] &= E[U|V,Z] \\ E[V|Z] &= 0, \end{aligned} \tag{10}$$

where X is a vector of endogeneous variables, Z is a vector of instruments, Π_0 is an unkown function, and V is a vector of disturbances. Then

$$\begin{aligned} E[Y|X,Z] &= g_0(X) + E[U|X,Z] \\ &= g_0(X) + E[U|V,Z] \\ &= g_0(X) + E[U|V] \\ &= g_0(X) + \nu(V) \\ \Rightarrow \quad Y &= g_0(X) + \nu(V) + \varepsilon, \quad E[\varepsilon|X,V] = 0 \end{aligned} \tag{11}$$

Notice that $\Pi_0(Z) = E[X|Z]$ is identified so that V is identified.

Nonparametric control function (Contd.)

Newey et al. (1999) provide various sufficient assumptions under which g_0 is identified. These conditions ensure that V and X are sufficiently distinct.

Theorem 2.1 of Newey et al. (1999): g_0 is identified, up to an additive constant, if and only if $P(\delta(X) + \gamma(V) = 0) = 1$ implies there exists a constant c_g such that $P(\delta(X) = c_g) = 1$.

For intuition, suppose that identification fails. Then, by the theorem, there are functions $\delta(X)$ and $\gamma(V)$ such that $\delta(X) + \gamma(V) = 0$ and $\delta(X)$ nonconstant. This implies a degeneracy in the joint distribution of these two random variables. Absence of such an exact relationship will imply identification. A sufficient condition is given by the following theorem:

Theroem 2.3 of Newey et al. (1999): If g_0, ν, Π_0 are differentiable, the boundary of the support of (Z, V) has zero probability, and $\Pi_0(Z) = \dim X$, then g_0 is identified.

This is a nonparametric generalization of the rank condition in TSLS.

Estimation in Newey et al. (1999)

The estimation of the Newey et al. (1999) is rather straightforward:

1: Required input: (Y, X, Z) the data, L the order of basis expansion in the first stage, (K_X, K_V) the order of basis expansions in the second stage;

2: **procedure**

3: $Z_L \leftarrow \text{BasisExpand}(Z, L)$ ▷ Obtain basis functions of Z

4: $\hat{\theta}_L \leftarrow (Z_L^\top Z_L)^{-1} Z_L^\top X$ ▷ Estimate Π_0

5: $\hat{V} \leftarrow X - \hat{\theta}_L^\top Z_L$

6: $X_K \leftarrow \text{BasisExpand}(X, K_X)$ ▷ Obtain basis functions of X

7: $\hat{V}_K \leftarrow \text{BasisExpand}(\hat{V}, K_V)$ ▷ Obtain basis functions of $\hat{V}$

8: $\tilde{X}_K \leftarrow [X_K, \hat{V}_K]^\top$

9: $[\hat{\theta}_g, \hat{\theta}_\nu] \leftarrow (\tilde{X}_K^\top \tilde{X}_K)^{-1} \tilde{X}_K^\top Y$

10: Return: $\hat{\theta}_g, \hat{\theta}_\nu$

We can thus compute $\hat{g}(x_2) - \hat{g}(x_1) = \hat{\theta}_g^\top (x_{2K} - x_{1K})$, where x_2 and x_1 are values of interest (e.g., status quo and policy proposal) and the subscript K denotes the corresponding basis expansion.

Notice that stability depends on the singular values of $(\tilde{X}_K^\top \tilde{X}_K)$.

References

Amemiya, T. (1974). Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica*, pages 999–1012.

Andrews, D. W. (2017). Examples of l2-complete and boundedly-complete distributions. *Journal of econometrics*, 199(2):213–220.

Blundell, R. and Powell, J. L. (2006). Endogeneity in nonparametric and semiparametric regression models. *Advances in Economics and Econometrics*, pages 312–357.

Canay, I. A., Santos, A., and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559.

Das, M. (2005). Instrumental variables estimators of nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, 124(2):335–361.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR.

Lehman, E. L. (1959). *Testing statistical hypotheses*. Wiley, New York.

Munkres, J. R. (2000). Topology.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Newey, W. K., Powell, J. L., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603.

Theil, H. (1953). Repeated least squares applied to complete equation systems. *The Hague: Central Planning Bureau*.

Notes on Completeness

back to main slides

Definition: A family $\mathcal{P}$ of probability distribution P is *complete* if

$$E_P[f(X)] = 0, \forall P \in \mathcal{P} \quad \Rightarrow \quad f(x) \overset{a.e.}{=} 0. \tag{12}$$

Theorem 1 in Lehman (1959): Let X be a random vector with probability distribution

$$dP_\theta(x) = C(\theta) \exp\left[\sum_{j=1}^{k} \theta_j T_j(x)\right] d\mu(x), \tag{13}$$

and let $\mathcal{P}^{\mathcal{T}}$ be the family of distributions of $T = (T_1(X), \ldots, T_k(X))$ as θ ranges over the set ω. Then $\mathcal{P}^{\mathcal{T}}$ is complete provided ω contains a k-dimensional rectangle.