cemmap
centre for microdata methods and practice

Random coefficients on endogenous variables in simultaneous equations models

Matthew Masten

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP25/15

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Random Coefficients on Endogenous Variables in Simultaneous Equations Models*

Matthew A. Masten
Department of Economics
Duke University
matt.masten@duke.edu

May 19, 2015

Abstract

This paper considers a classical linear simultaneous equations model with random coefficients on the endogenous variables. Simultaneous equations models are used to study social interactions, strategic interactions between firms, and market equilibrium. Random coefficient models allow for heterogeneous marginal effects. I show that random coefficient seemingly unrelated regression models with common regressors are not point identified, which implies random coefficient simultaneous equations models are not point identified. Important features of these models, however, can be identified. For two-equation systems, I give two sets of sufficient conditions for point identification of the coefficients' marginal distributions conditional on exogenous covariates. The first allows for small support continuous instruments under tail restrictions on the distributions of unobservables which are necessary for point identification. The second requires full support instruments, but allows for nearly arbitrary distributions of unobservables. I discuss how to generalize these results to many equation systems, where I focus on linear-in-means models with heterogeneous endogenous social interaction effects. I give sufficient conditions for point identification of the distributions of these endogenous social effects. I suggest a nonparametric kernel estimator for these distributions based on the identification arguments. I apply my results to the Add Health data to analyze peer effects in education.

*This is a revised version of my Nov 3, 2012 job market paper. I am very grateful for my advisor, Chuck Manski, for his extensive support and encouragement. I am also grateful for my committee members, Ivan Canay and Elie Tamer, who have been generous with their advice and feedback. I also thank Federico Bugni, Mark Chicu, Joachim Freyberger, Jeremy Fox, Jin Hahn, Stefan Hoderlein, Joel Horowitz, Shakeeb Khan, Rosa Matzkin, Konrad Menzel, Alex Torgovitsky, and Daniel Wilhelm for helpful discussions and comments, and seminar participants at Northwestern University, UCLA, University of Pittsburgh, Duke University, University of Chicago Booth School of Business, Federal Reserve Board of Governors, Midwest Economics Association Annual Meeting, the CEME Stanford/UCLA Conference, Boston College, University of Iowa, University of Tokyo, Princeton University, and Columbia Unviersity. I thank Margaux Luflade for excellent research assistance. This research was partially supported by a research grant from the University Research Grants Committee at Northwestern University. This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris; see references for full citation and acknowledgements. This research uses data from the AHAA study; see references for full citation and acknowledgements.

1 Introduction

Simultaneous equations models are among the oldest models studied in econometrics. Their importance arises from economists' interest in equilibrium situations, like social interactions, strategic interactions between firms, and market equilibrium. They are also the foundation of work on treatment effects and self-selection. The classical linear simultaneous equations model assumes constant coefficients, which implies that all marginal effects are also constant. While there has been much work on allowing for heterogeneous marginal effects by introducing random coefficients on exogenous variables, or on endogenous variables in triangular systems, there has been little work on random coefficients on endogenous variables in fully simultaneous systems. In this paper, I consider identification and estimation in such systems. For example, I provide sufficient conditions for point identification of the distribution of elasticities across markets in a simple supply and demand model with linear equations.

I consider the system of two linear simultaneous equations

$$\begin{aligned} Y_1 &= \gamma_1 Y_2 + \beta_1 Z_1 + \delta_1' X + U_1 \\ Y_2 &= \gamma_2 Y_1 + \beta_2 Z_2 + \delta_2' X + U_2, \end{aligned} \tag{1}$$

where $Y \equiv (Y_1, Y_2)'$ are observable outcomes of interest which are determined simultaneously as the solution to the system, $Z \equiv (Z_1, Z_2)'$ are observable instruments, X is a K-vector of observable covariates, and $U \equiv (U_1, U_2)'$ are unobservable variables. X may include a constant. Note that, while important for applied work, the covariates X will play no role in the identification arguments; see remark 2 on page 23. In the data, we observe the joint distribution of (Y, Z, X). This system is triangular if one of γ_1 or γ_2 is known to be zero; it is fully simultaneous otherwise. Two exclusion restrictions are imposed: Z_1 only affects Y_1, and Z_2 only affects Y_2. These exclusion restrictions, plus the assumption that Z and X are uncorrelated with U, can be used to point identify $(\gamma_1, \gamma_2, \beta_1, \beta_2, \delta_1, \delta_2)$, assuming these coefficients are all constants.[1]

I relax the constant coefficient assumption by allowing γ_1 and γ_2 to be random. The distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$, or features of these distributions like the means $\mathbb{E}(\gamma_1 \mid X)$ and $\mathbb{E}(\gamma_2 \mid X)$, are the main objects of interest. For example, we may ask how the average effect of Y_2 on Y_1 changes if we increase a particular covariate. Classical mean-based identification analysis may fail with random γ_1 and γ_2 due to non-existence of reduced form mean regressions. Moreover, I show that random coefficient seemingly unrelated regression models with common regressors are not point identified, which implies that random coefficient simultaneous equations models are not point identified. Despite this, I prove that the marginal distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$ are point identified if the instruments Z have full support and are independent of all unobservables. I show

[1] This result, along with further discussion of the classical model with constant coefficients, is reviewed in most textbooks. Also see the handbook chapters of Hsiao (1983), Intriligator (1983), and Hausman (1983), as well as the classic book by Fisher (1966). Model (1) applies to continuous outcomes. For simultaneous systems with discrete outcomes, see Bjorn and Vuong (1984), Bresnahan and Reiss (1991), and Tamer (2003).

that, with tail restrictions on the distribution of unobservables, full support Z can be relaxed. I show that these tail restrictions are necessary for point identification when Z has bounded support. I propose a consistent nonparametric estimator for the distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$.

I then show how to extend the identification arguments to systems with more than two equations. A general linear system of N simultaneous equations with random coefficients has $O(N^2)$ coefficients, compared to the $2N$ dimensional distribution of outcomes and instruments. This dimensionality problem implies that it is generally not possible to identify the entire joint distribution of all these coefficients. Nonetheless, under restrictions that reduce the dimensionality of the random coefficients, we can recover point identification. While there are many possible restrictions one could consider, I focus on a random coefficients generalization of the most widely used social interactions model—the linear-in-means model (Manski 1993). Specifically, I consider the model

$$Y_i = \gamma_i \frac{1}{N-1} \sum_{j \neq i} Y_j + \beta_i Z_i + \delta_i' X_i + U_i. \tag{2}$$

Here person i's outcome depends on the average of the other $N-1$ people in their reference group. γ_i is called the *endogenous social interaction parameter*. The classical linear-in-means model assumes γ_i is constant across all people i, while the random coefficients linear-in-means model allows it to vary across individuals. I also consider a generalization which incorporates observed network data. In both cases I give sufficient conditions for point identification of the distribution of the endogenous social interaction parameter. These conditions are similar to those in the two equation case.

Throughout I assume all coefficients on exogenous variables are also random. Note that the additive unobservables can be thought of as random coefficients on a constant covariate. Throughout the paper, I use the following application as a leading example of a two-equation system (also see Moffitt 2001).

Example (Social interactions between pairs of people)**.** *Consider a population of pairs of people, such as spouses, siblings, or best friends. Let Y_1 denote the outcome for the first person and Y_2 the outcome for the second. These outcomes may be hours worked, GPA, body weight, consumption, savings, investment, etc. Model* (1) *allows for endogenous social interactions: one person's outcome may affect the other person's, and vice versa. Because I allow for random coefficients, these social interaction effects are not required to be constant across all pairs of people.*

Social interaction models for household behavior have a long history within labor and family economics (see Browning, Chiappori, and Weiss 2014 for a survey). Recently, several papers have studied social interactions between 'ego and alter' pairs of people, or between pairs of 'best friends', studying outcomes like sexual activity (Card and Giuliano 2013), obesity (Christakis and Fowler 2007, Cohen-Cole and Fletcher 2008), and educational achievement (Sacerdote 2001). In an empirical application, I study peer effects in educational achievement. I use the Add Health data to

construct best friend pairs. I set the outcomes Y_1 and Y_2 to be each friends' GPA, and following one specification in Sacerdote (2000, 2001) I choose Z_1 and Z_2 to be each friends' lagged GPA. I then estimate the distributions of γ_1 and γ_2 and find evidence for substantial heterogeneity in social interaction effects and that usual point estimates are smaller than the nonparametrically estimated average social interaction effect.

In the rest of this section, I review the related literature. Kelejian (1974) and Hahn (2001) are the only papers explicitly about random coefficients on endogenous variables in simultaneous systems. Kelejian considers a linear system like (1) and derives conditions under which we can apply traditional arguments based on reduced form mean regressions to point identify the means of the coefficients. These conditions rule out fully simultaneous systems. For example, with two equations they imply that the system is triangular (see remark 3 on page 50). Furthermore, Kelejian assumes all random coefficients are independent of each other, which I do not require. Hahn considers a linear simultaneous equations model like system (1). He applies a result of Beran and Millar (1994) which requires the joint support of all covariates across all reduced form equations to contain an open ball. This is not possible in the reduced form for system (1) since each instrument enters more than one reduced form equation (see remark 4 on page 50).

Random coefficients on exogenous variables, in contrast, are well understood. The earliest work goes back to Rubin (1950), Hildreth and Houck (1968), and Swamy (1968, 1970), who propose estimators for the mean of a random coefficient in single equation models. See Raj and Ullah (1981, page 9) and Hsiao and Pesaran (2008) for further references and discussion. More recent work has focused on estimating the distribution of random coefficients (Beran and Hall 1992, Beran and Millar 1994, Beran 1995, Beran, Feuerverger, and Hall 1996, and Hoderlein, Klemelä, and Mammen 2010).

Random coefficients on endogenous variables in triangular systems are also well studied (Heckman and Vytlacil 1998, Wooldridge 1997, 2003). For example, suppose $\gamma_2 \equiv 0$ and γ_1 is random. If β_2 is constant then $\mathbb{E}(\gamma_1)$ is point identified and can be estimated by 2SLS. If β_2 is random, then the 2SLS estimand is a weighted average of γ_1—a parameter similar to the weighted average of local average treatment effects (Angrist and Imbens 1995). This model has led to a large literature on instrumental variables methods with heterogeneous treatment effects; that is, generalizations of a linear model with random coefficients on an endogenous variable (Angrist 2004).

For discrete outcomes, random coefficients have been studied in many settings. Ichimura and Thompson (1998), Fox, Kim, Ryan, and Bajari (2012), and Gautier and Kitamura (2013) study binary outcome models with exogenous regressors. Gautier and Hoderlein (2012) and Hoderlein and Sherman (2013) study triangular systems. Finally, recent work by Dunker, Hoderlein, and Kaido (2013) and Fox and Lazzati (2013) study random coefficients in discrete games.

A large recent literature has examined nonseparable error models like $Y_1 = m(Y_2, U_1)$, where m is an unknown function (e.g. Matzkin 2003, Chernozhukov and Hansen 2005, and Torgovitsky 2014). These models provide an alternative approach to allowing heterogeneous marginal effects. Although

many papers in this literature allow for Y_2 to be correlated with U_1, they typically assume that U_1 is a scalar, which rules out models with both an additive unobservable and random coefficients, such as the first equation of system (1). Additionally, m is typically assumed to be monotonic in U_1, which imposes a rank invariance restriction. For example, in supply and demand models, rank invariance implies that the demand functions for any two markets cannot cross. The random coefficient system (1) allows for such crossings. A related literature on nonlinear and nonparametric simultaneous equations models also allows for nonseparable errors (see Brown 1983, Roehrig 1988, Benkard and Berry 2006, Matzkin 2008, Blundell and Matzkin 2014, and Berry and Haile 2011, 2014), but these papers again restrict the dimension of unobservables by assuming that the number of unobservables equals the number of endogenous variables.

Several papers allow for both nonseparable errors and vector unobservables U_1, but make assumptions which rule out model (1) with random γ_1 and γ_2. Imbens and Newey (2009) and Chesher (2003, 2009) allow for a vector unobservable, but restrict attention to triangular structural equations. Hoderlein and Mammen (2007) allow for a vector unobservable, but require independence between the unobservable and the covariate (i.e., $Y_2 \perp\!\!\!\perp U_1$ in the above model), which cannot hold in a simultaneous equations model.

Finally, several papers allow for both simultaneity and high dimensional unobservables. Matzkin (2012) considers a simultaneous equations model with more unobservables than endogenous variables, but assumes that the endogenous variables and the unobservables are additively separable. Fox and Gandhi (2011) consider a nonparametric system of equations with nonadditive unobservables of arbitrary dimension. They assume all unobservables have countable support, which implies that outcomes are discretely distributed, conditional on covariates. I focus on continuously distributed outcomes. Angrist, Graddy, and Imbens (2000) examine the two equation supply and demand example without imposing linearity or additive separability of a scalar unobserved heterogeneity term. Following their work on LATE, they show that with a binary instrument the traditional linear IV estimator of the demand slope converges to a weighted average of the average derivative of the demand function over a subset of prices. Their assumptions are tailored to the supply and demand example and they do not consider identification of the distribution of marginal effects. Manski (1995, 1997) considers a general model of treatment response. Using a monotonicity assumption, he derives bounds on observation level treatment response functions. These bounds hold regardless of how treatment is selected and thus apply to simultaneous equations models. He shows how these observation level bounds imply bounds on parameters like average demand functions. I impose additional structure which allows me to obtain stronger identification results. I also do not require monotonicity. Okumura (2011) builds on the monotonicity based bounds analysis of Manski, deriving bounds on the medians and cdfs of the unobservables in a simultaneous equations model with nonparametric supply and demand functions which each depend on a scalar unobservable. Kasy (2014) studies general nonparametric systems with arbitrary dimensional unobservables, but focuses attention on identifying average structural functions via a monotonicity

condition. Hoderlein, Nesheim, and Simoni (2012) study identification and estimation of distributions of unobservables in structural models. They assume that a particular scalar unobservable has a known distribution, which I do not require. They also focus on point identification of the entire distribution of unobservables, which in system (1) includes the additive unobservables and the coefficients on exogenous variables. As I mentioned above, the entire joint distribution of unobservables in (1) is not point identified, and hence I focus on identification of the distribution of endogenous variable coefficients only.

2 The simultaneous equations model

Consider again system (1), the linear simultaneous equations model:

$$\begin{aligned} Y_1 &= \gamma_1 Y_2 + \beta_1 Z_1 + \delta_1' X + U_1 \\ Y_2 &= \gamma_2 Y_1 + \beta_2 Z_2 + \delta_2' X + U_2. \end{aligned} \tag{1}$$

Assume β_1 and β_2 are random scalars, δ_1 and δ_2 are random K-vectors, and γ_1 and γ_1 are random scalars. In matrix notation, system (1) is

$$Y = \Gamma Y + BZ + DX + U,$$

where

$$\Gamma = \begin{pmatrix} 0 & \gamma_1 \\ \gamma_2 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix}, \quad \text{and} \quad D = \begin{pmatrix} \delta_1' \\ \delta_2' \end{pmatrix}.$$

Let I denote the identity matrix. When $(I - \Gamma)$ is invertible (see section 2.1 below), we can obtain the reduced form system

$$Y = (I - \Gamma)^{-1} BZ + (I - \Gamma)^{-1} DX + (I - \Gamma)^{-1} U.$$

Writing out both equations in full yields

$$\begin{aligned} Y_1 &= \frac{1}{1 - \gamma_1 \gamma_2} \left[U_1 + \gamma_1 U_2 + \beta_1 Z_1 + \gamma_1 \beta_2 Z_2 + \delta_1' X + \gamma_1 \delta_2' X \right] \\ Y_2 &= \frac{1}{1 - \gamma_1 \gamma_2} \left[\gamma_2 U_1 + U_2 + \gamma_2 \beta_1 Z_1 + \beta_2 Z_2 + \gamma_2 \delta_1' X + \delta_2' X \right]. \end{aligned} \tag{3}$$

Identification follows from examining this reduced form system.

Depending on the specific empirical application, the signs of γ_1 and γ_2 may both be positive, both be negative, or have opposite signs. When analyzing social interactions between pairs of people, like spouses or best friends, we expect positive, reinforcing social interaction effects; both γ_1 and γ_2 are positive. If we analyze strategic interaction between two firms, such as in the classical Cournot duopoly model, we expect negative interaction effects; both γ_1 and γ_2 are negative. In the

classical supply and demand model, supply slopes up and demand slopes down; the slopes γ_1 and γ_2 have opposite signs.

2.1 Unique solution

For a fixed value of (Z, X), there are three possible configurations of system (1), depending on the realization of (B, D, U, Γ): parallel and overlapping lines, parallel and nonoverlapping lines, and non-parallel lines. Figure 1 plots each of these configurations.

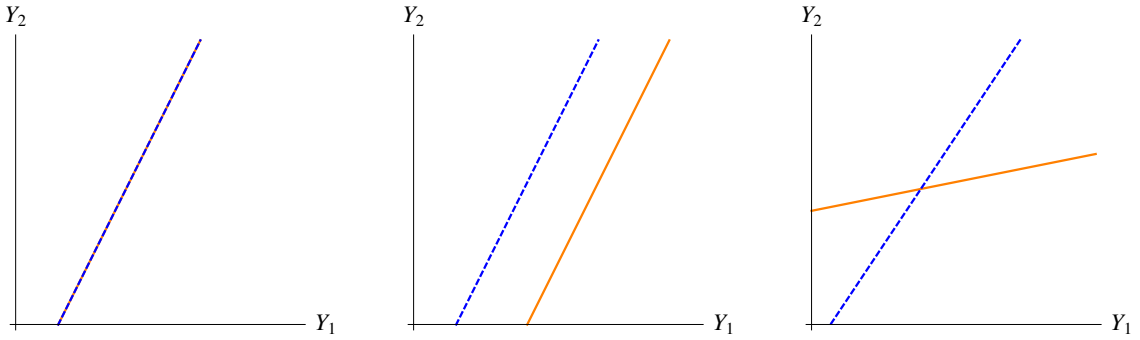


Figure 1: These figures plot the lines $Y_1 = \gamma_1 Y_2 + C_1$, shown as the solid line, and $Y_2 = \gamma_2 Y_1 + C_2$, shown as the dashed line. By varying γ_1, γ_2, C_1, and C_2, each plot shows a different possible configuration of the system: parallel and overlapping, parallel and nonoverlapping, and non-parallel.

When (B, D, U, Γ) are such that the system has non-parallel lines, the model specifies that the observed outcome Y is the unique solution to system (1). In the case of parallel and overlapping lines, the model specifies that the observed outcome Y lies on that line, but it does not predict a unique Y. Finally, when the system has parallel and nonoverlapping lines, the model makes no prediction and the observed Y is generated from some unknown distribution. Because of these last two cases, the model is incoherent and incomplete without further assumptions (see Tamer 2003 and Lewbel 2007 for a discussion of coherency and completeness). To ensure coherency and completeness, I make the following assumption, which implies that a unique solution to system (1) exists with probability 1.[2]

Assumption A1 (Existence of a unique solution). $\mathbb{P}(\gamma_1\gamma_2 = 1 \mid X, Z) = 0$.

Since $\det(I - \Gamma) = 1 - \gamma_1\gamma_2$, this assumption is equivalent to requiring $(I - \Gamma)$ to be invertible with probability 1 (conditional on X, Z), which allows us to work with the reduced form system (3). A1 rules out the first two configurations of system (1) almost surely, since parallel lines occur when $\gamma_1 = 1/\gamma_2$, or equivalently when $\gamma_1\gamma_2 = 1$. The existing literature on simultaneous equations with continuous outcomes, including both classical linear models with constant coefficients as well

[2]Here and throughout the paper, stating that an assumption which holds 'given X' means that it holds given $X = x$ for all $x \in \text{supp}(X)$, where $\text{supp}(X)$ denotes the support of X. This can be relaxed to hold only at x values for which we wish to identify the distribution of $\gamma_i \mid X = x$, $i = 1, 2$, or to hold only X-almost everywhere if we are only interested in the unconditional distribution of γ_i.

as recent nonparametric models, makes a unique solution assumption analogous to A1. Indeed, in the linear model (1) with constant coefficients, relaxing the unique solution assumption implies that $\gamma_1\gamma_2 = 1$ in every system. Hence only the two parallel line configurations may occur. In that case, it is possible that the distribution of (U_1, U_2) is such that the lines never overlap, which implies that constant coefficient model with $\gamma_1\gamma_2 = 1$ places no restrictions on the data.

When (γ_1, γ_2) are random coefficients, there is scope for relaxing A1 without obtaining a vacuous model, although I do not pursue this in depth. For example, we could replace A1 with the assumption $\mathbb{P}(\gamma_1\gamma_2 = 1 \mid X, Z) < p$ for some known p, $0 \leq p < 1$. This says that the model delivers a unique outcome in $100(1-p)$ percent of the systems. In the remaining systems, the model does not. Thus, even if we are unwilling to make assumptions about how the outcome data Y are generated when $\gamma_1\gamma_2 = 1$, we may still be able to obtain useful partial identification results, since we know that a unique solution occurs with at least probability p. This approach is similar to analysis of contaminated data (see Horowitz and Manski 1995).

2.2 Nearly parallel lines and fat tailed distributions

Although A1 rules out exactly parallel lines, it allows for *nearly* parallel lines. Nearly parallel lines occur when $\gamma_1\gamma_2$ is close, but not equal, to 1. In this case, $1 - \gamma_1\gamma_2$ is close to zero, and thus $1/(1-\gamma_1\gamma_2)$ is very large. This is problematic since $1/(1-\gamma_1\gamma_2)$ appears in all terms in the reduced form system (3). So, if $\gamma_1\gamma_2$ is close to 1 with high enough probability, the means of the random coefficients in the reduced form do not exist. This possibility precludes the classical mean-based identification approach of examining $\mathbb{E}(Y_1 \mid X, Z)$ and $\mathbb{E}(Y_2 \mid X, Z)$, without further restrictions on the distribution of (γ_1, γ_2).

In section 3, I show that even when these means fail to exist, we can still identify the marginal distributions of γ_1 and γ_2, under the assumption that Z has full support. I then replace full support Z with the weaker assumption that Z has continuous variation. The trade-off for this change is that I restrict the distribution of (γ_1, γ_2) by assuming that the reduced form coefficients *do not* have fat tails, so that their means do exist. Thus, in order to relax full support, I eliminate near parallel lines.

Remark 1. A similar mean non-existence issue arises in Graham and Powell's (2012) work on panel data identification of single equation correlated random coefficient models. Since their denominator term (see equation 22) is an observable random variable, they are able to use trimming to solve the problem. Here the denominator is unobserved and so we do not see which observations in the data are problematic. Hence I take a different approach. □

2.3 Two-stage least squares

As just discussed, nearly parallel lines can preclude mean-based identification approaches. In this case, the reduced form mean regressions $\mathbb{E}(Y_1 \mid X, Z)$ and $\mathbb{E}(Y_2 \mid X, Z)$ may not exist, and hence

any estimate of them, such as OLS of Y_1 and Y_2 on (X, Z), may fail to converge. Likewise, the 2SLS estimand may not exist, and so the 2SLS estimator also may fail to converge. Even when these means do exist, 2SLS will converge to a weighted average effect parameter, as shown by Angrist et al. (2000). To see this in the context of the linear model (1), suppose we are only interested in the first structural equation. Combining the structural equation for Y_1 (the first equation of system 1) with the reduced form equation for Y_2 (the second equation of system 3) yields

$$\begin{aligned} Y_1 &= \gamma_1 Y_2 + U_1 \\ Y_2 &= \pi_{21} + \pi_{23} Z_2, \end{aligned}$$

where I let $\delta_1 = \delta_2 = \beta_1 = 0$ for simplicity, and denote

$$\pi_2 = (\pi_{21}, \pi_{23}) = \left(\frac{U_2 + \gamma_2 U_1}{1 - \gamma_1 \gamma_2}, \frac{\beta_2}{1 - \gamma_1 \gamma_2} \right).$$

This is a triangular system of equations where γ_1 and π_2 are random and Z_2 is an instrument for Y_2. Let $\widehat{\gamma}_1$ denote the 2SLS estimator of γ_1. Assuming the relevant means exist (see section 2.2), we have

$$\widehat{\gamma}_1 \xrightarrow{p} \frac{\text{cov}(Y_1, Z_2)}{\text{cov}(Y_2, Z_2)} = \mathbb{E}\left[\frac{\beta_2/(1 - \gamma_1\gamma_2)}{\mathbb{E}[\beta_2/(1 - \gamma_1\gamma_2)]} \gamma_1 \right].$$

Thus 2SLS converges to a weighted average effect parameter (see appendix A for the derivations). This occurs even if β_2 is constant and therefore cancels out in the above expression. With constant β_2, if γ_2 is degenerate on zero, so that the system is not actually simultaneous, then 2SLS recovers $\mathbb{E}(\gamma_1)$, the mean random coefficient. The 2SLS estimand is commonly interpreted as weighting treatment effects by the heterogeneous instrument effect. Here, even when β_2 is a constant so that the instrument has the same effect on all people, heterogeneous effects of endogenous variables combined with simultaneity cause 2SLS to estimate a weighted average effect parameter. Observations in systems which are close to having parallel lines count the most. In this paper, I give conditions under which we can go beyond this weighted average effect parameter and identify the entire marginal distribution of each random coefficient.

3 Identification

In this section I study identification of random coefficients models. I first discuss two sets of sufficient conditions for point-identification in single equation random coefficients models. I later apply these results to simultaneous equation systems. I then discuss seemingly unrelated regressions. I show that when the equations have common regressors, the joint distribution of random coefficients is not point identified. This implies that simultaneous equations models with random coefficients are not point identified. I then move on to the simultaneous two equation system, where I show that despite the overall lack of point identification, we can point identify the marginal distributions of

endogenous variable coefficients. I discuss a special case where we can identify the joint distribution of these endogenous variable coefficients. I also consider identification in triangular models. Finally, I end with a discussion of the many equation case, where I give two results for linear-in-means social interaction models.

Throughout this paper, 'identified' means 'point identified'. See Matzkin (2007) for the formal definition of identification. Relaxing my sufficient conditions may lead to useful partial identification results for the features of interest. Since such partial identification results have not been explored even in single equation random coefficient models, I leave this to future research.

3.1 Single equation models

In this section I discuss two lemmas about identification of single equation random coefficient models. These lemmas are important steps in the proofs of theorems 2 and 3 on simultaneous equations models. The first lemma allows for arbitrary distributions of random coefficients, but requires full support regressors.

Lemma 1. Suppose

$$Y = A + B'Z,$$

where Y and A are scalar random variables and B and Z are random K-dimensional vectors. Suppose the joint distribution of (Y, Z) is observed. If $Z \perp\!\!\!\perp (A, B)$ and Z has support $\mathbb{R}^K$ then the joint distribution of (A, B) is identified.

While I gather all proofs in appendix A, I sketch the proofs here to show their main ideas. The proof of this lemma is similar to that of the classical Cramér-Wold theorem (Cramér and Wold 1936 page 291; see also Beran and Millar 1994 page 1980 and Ichimura and Thompson 1998 theorem 1) that the joint distribution of a random vector is uniquely determined by its one-dimensional projections. The proof follows by examining the characteristic function of Y given Z:

$$\begin{aligned}\phi_{Y|Z}(t \mid z_1, \ldots, z_K) &= \mathbb{E}[\exp(it(A + B_1 Z_1 + \cdots + B_K Z_K)) \mid Z = (z_1, \ldots, z_K)] \\ &= \phi_{A,B}(t, tz_1, \ldots, tz_K),\end{aligned}$$

where the second line follows since $Z \perp\!\!\!\perp (A, B)$ and by the definition of the characteristic function for (A, B). Here B_k is the scalar random coefficient on Z_k. Thus, by varying $(z_1, \ldots, z_K)$ over $\mathbb{R}^K$, and t over $\mathbb{R}$, we can learn the entire characteristic function of (A, B).

The following result relaxes the full support condition, but imposes tail conditions on the distribution of random coefficients. When Z has bounded support, these tail conditions are necessary if we wish to obtain point identification.

Lemma 2. Suppose

$$Y = A + B'Z,$$

where Y and A are scalar random variables and B and Z are random K-dimensional vectors. Suppose the joint distribution of (Y, Z) is observed. Assume

1. $Z \perp\!\!\!\perp (A, B)$ and
2. $\text{supp}(Z)$ contains an open ball in $\mathbb{R}^K$.

Then

3. the distribution of (A, B) has finite absolute moments, and
4. the distribution of (A, B) is uniquely determined by its moments

are sufficient for identification of the joint distribution of (A, B). If $\text{supp}(Z)$ is bounded, then (3) and (4) are also necessary for identification of the joint distribution of (A, B), as well as identification of each marginal distribution of regressor coefficients B_k, $k = 1, \ldots, K$. If (4) does not hold, then distribution of the intercept A is point identified if and only if $0 \in \text{supp}(Z)$.

For a scalar Z, the sufficiency direction of this result was proved in Beran's (1995) proposition 2. Lemma 2 here shows that the sufficiency result holds for any finite dimensional vector Z, as used in the simultaneous equations analysis, uses a different proof technique, and also shows the necessity direction. The proof of sufficiency is a close adaptation of the proofs of theorem 3.1 and corollary 3.2 in Cuesta-Albertos, Fraiman, and Ransford (2007), who prove a version of the classical Cramér-Wold theorem. I first show that all moments of (A, B) are identified, and then conclude that the distribution is identified from its moments. Because of this proof strategy, if we are only interested in moments of (A, B) in the first place—say, the first and second moment—then we do not need assumption (4) in lemma 2.

The necessity direction follows by a counterexample given in Bélisle, Massé, and Ransford (1997). It exhibits two distributions which have the same projections onto lines defined by the support of Z (so they are observational equivalent in the random coefficient model), whose moments are all finite and equal, and yet actually have disjoint support. To get some intuition for this result, recall that analytic functions are uniquely determined by their values in any small neighborhood in their domain. So if the characteristic function of (A, B) is analytic, then 'local' variation of the regressors is sufficient to know this characteristic function in a small neighborhood, and hence by analyticity is sufficient for knowing the entire characteristic function. Having an analytic characteristic function is a kind of thin tail condition. Roughly speaking, once the distribution has fatter tails, we loose analyticity, and then knowledge the characteristic function in a small neighborhood is not sufficient for knowledge of the entire function. The formal argument is more complicated because (A, B) having an analytic characteristic function is actually not necessary for point identification. But the idea is similar: achieving point identification of (A, B) with small support Z requires some kind of extrapolation. Lemma 2 shows that the weaker conditions (3) and (4) are precisely what is necessary.

In this counterexample, assumption (3) holds while assumption (4) fails. Moreover, although for clarity I have listed assumptions (3) and (4) separately, assumption (4) actually implies assumption (3); see lemma 5 in the appendix. Hence if (4) is necessary, then so is (3). Finally, the necessity direction also shows that even the marginal distributions of regressor coefficients are not point identified if (4) fails. This final step will be important for applying this necessity result to simultaneous equations models.

This necessity result depends on allowing all the coefficients to be random. In the textbook constant coefficients linear model, it is well known that point identification of the constant coefficients and the distribution of the random intercept is possible even if the random intercept is Cauchy distributed. This classical result uses the assumption of constant coefficients and hence does not contradict lemma 2.

3.2 Seemingly unrelated regressions

A seemingly unrelated regressions (SUR) model is a system of equations

$$\begin{aligned} Y_1 &= A_1 + B_1' Z_1 \\ &\vdots \\ Y_N &= A_N + B_N' Z_N \end{aligned} \tag{4}$$

which are related in that their unobservable terms are correlated. Consequently, in the classical case where $A \equiv (A_1, \ldots, A_N)$ are random intercepts and $B \equiv (B_1, \ldots, B_N)$ are constant coefficients, a more efficient estimator of each coefficient B_n can be obtained by estimating all coefficients simultaneously. Moreover, sometimes cross equation constraints on the coefficients are imposed, like $B_1 = B_2$, which also implies that joint estimation will improve efficiency. Here the regressors Z_n are subscripted across equations, but this notation allows for common regressors.

In this section I consider the SUR model where $B \equiv (B_1, \ldots, B_N)$ are random coefficients. For simplicity I focus on the two equation case:

$$\begin{aligned} Y_1 &= A_1 + B_1' Z_1 \\ Y_2 &= A_2 + B_2' Z_2 \end{aligned} \tag{4'}$$

although the main result extends immediately to the general system (4).

Say there is a *functional dependence* of X on W if there exists a function $f : \mathbb{R} \to \mathbb{R}$ such that $X = f(W)$ almost surely. The following result strengthens proposition 2.2 of Beran and Millar (1994) by providing weaker moment conditions (they assumed the unobservables had compact support), as well as showing that, given the other assumptions, functional relationships between covariates preclude point identification.

Theorem 1. Consider the SUR system $4'$ where Y_1, Y_2, A_1, and A_2 are scalar random variables, B_1 and Z_1 are random K_1-dimensional vectors, and B_2 and Z_2 are random K_2-dimensional vectors. Suppose the joint distribution of (Y_1, Y_2, Z_1, Z_2) is observed. Assume

1. $(Z_1, Z_2) \perp\!\!\!\perp (A, B)$,
2. $\text{supp}(Z_1, Z_2)$ contains an open ball in $\mathbb{R}^{K_1+K_2}$,
3. the distribution of (A, B) has finite absolute moments, and
4. the distribution of (A, B) is uniquely determined by its moments.

Then the joint distribution of (A, B) is point identified. If there is a functional dependence of *any* component of (Z_1, Z_2) on any other component, in which case (2) fails, then the joint distribution of (A, B) is not point identified. If $\text{supp}(Z)$ is bounded, then (3) and (4) are necessary for identification of the joint distribution of (A, B), as well as identification of each marginal distribution of regressor coefficients.

Necessity of the moments assumption follows because it is necessary in single equation models. Next, suppose X_1 and X_2 are functionally related components of (Z_1, Z_2). Then the distribution of X_1 conditional on X_2 is degenerate. We cannot independently vary X_1 and X_2 in the data. This means that we only know the characteristic function of the random coefficients on a linear subspace of $\mathbb{R}^{K_1+K_2+2}$. From the theory of characteristic functions and Fourier transforms, knowledge of a Fourier transform on such a subspace, which has measure zero, is not sufficient to pin down the original function (Cuesta-Albertos et al. 2007).

Corollary 1. Under the assumptions of theorem 1, if Z_1 and Z_2 contain a common regressor, then the joint distribution of (A_1, A_2, B_1, B_2) is not point identified.

As discussed in the next section, this corollary implies that the joint distribution of random coefficients in the linear simultaneous equations model is necessarily not point identified. In the SUR model, the joint distribution of coefficients (A_n, B_n) from any single equation n is point identified by applying the single equation lemmas 1 or 2. The result in this corollary is that the full joint distribution of all cross-equation coefficients is not point identified. Intuitively, consider the two-equation model with a single scalar covariate Z which is common to both equations:

$$Y_1 = A_1 + B_1 Z$$
$$Y_2 = A_2 + B_2 Z.$$

When examining the distribution of $(Y_1, Y_2) \mid Z = z$, variation in z affects both equations simultaneously. We cannot independently vary the regressor in the first equation from the regressor in the second equation. Regardless of the support of Z, this implies that the characteristic function

of (A_1, A_2, B_1, B_2) is known only on a linear subspace of $\mathbb{R}^4$, which is not sufficient to pin down its distribution.

Another result that follows from the previous theorem is that nonlinear random coefficient models are not point identified.

Corollary 2. Consider the nonlinear single equation random coefficients model

$$Y = A + B' p^K(Z)$$

where Z is a scalar and

$$p^K(z) = (p_{1K}(z), \ldots, p_{KK}(z))'$$

is a vector of known basis functions. Assume $Z \perp\!\!\!\perp (A, B)$. Assume the distribution of (A, B) has finite absolute moments and is uniquely determined by its moments. Then the joint distribution of (A, B) is not point identified.

This result is similar to the fact that constant coefficients are not point identified in linear models if there is perfect multicollinearity (i.e., if the support of the regressors lies in a proper linear subspace). The difference here is that nonlinear transformations are not sufficient to break the non-identification result. The intuition for this result is similar to the problem with common regressors in the SUR model: our inability to vary two regressors independently precludes identification of the joint distribution of their coefficients.

3.3 Simultaneous equations models

In this section I prove several point-identification results for the simultaneous equations model (1). The two main results give sufficient conditions for point identification of the marginal distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$. I also provide results on identification of the joint distribution of $(\gamma_1, \gamma_2) \mid X$ as well as on identification in triangular models.

The first main result supposes the instruments Z have continuous variation, but allows them to have bounded support. As in single equation models (lemma 2), a moment determinacy condition on the distribution of unobservables is necessary for point identification. The second main result shows that this moment determinacy condition is not necessary if Z has unbounded support. In this case we are able to identify the marginal distributions $\gamma_1 \mid X$ and $\gamma_2 \mid X$ even if the reduced form mean regression fails to exist because the structural equations are nearly parallel too often.

In addition to the unique solution assumption A1, I place several other restrictions on the unobservables.

Assumption A2 (Relevance). $\mathbb{P}(\beta_1 = 0 \mid X) = 0$ and $\mathbb{P}(\beta_2 = 0 \mid X) = 0$.

For units with $\beta_1 = 0$, given A3 below, Z_1 has no effect whatsoever on the distribution of $(Y_1, Y_2) \mid X$ and hence cannot help with identification; likewise for units with $\beta_2 = 0$. This

difficulty of learning causal effects for units whom are not affected by the instrument is well known and is not particular to the model considered here. As in the existing literature, such as the work on LATE, A2 can be relaxed if we only wish to identify causal effects for the subpopulation of units whom are affected by the instrument. That is, if $\mathbb{P}(\beta_1 = 0 \mid X) > 0$, then we can identify the distribution of γ_2 conditional on X and $\beta_1 \neq 0$. Likewise, if $\mathbb{P}(\beta_2 = 0 \mid X) > 0$, then we can identify the distribution of γ_1 conditional on X and $\beta_2 \neq 0$. Moreover, as in the constant coefficients case, if we are only interested in one equation, then we do not need an instrument for the other equation. That is, $\mathbb{P}(\beta_1 = 0 \mid X) > 0$ is allowed if we only wish to identify the distribution of $\gamma_1 \mid X$. If we only wish to identify the distribution of $\gamma_2 \mid X$, then $\mathbb{P}(\beta_2 = 0 \mid X) > 0$ is allowed.

Assumption A3 (Independence). $Z \perp\!\!\!\perp (B, D, U, \Gamma) \mid X$.

Nearly all of the literature on random coefficients models with cross-sectional data makes an independence assumption similar to A3.[3] Moreover, this assumption commonly is maintained throughout the literature on nonparametric nonseparable models, in single equation, triangular, and simultaneous equations models.[4] See Berry and Haile (2014) for further discussion of instruments often used in simultaneous equations models, along with extensive citations to empirical research.

This assumption reduces the complexity of the model by restricting how the distribution of unobservables can depend on the observed covariates: the distribution of (B, D, U, Γ) is assumed to be the same regardless of the realization of Z, conditional on X. The covariates X may still be correlated with the unobservables, and (Y_1, Y_2), as outcome variables, are generally also correlated with all of the unobservables.

Example (Social interactions between pairs of people, cont'd)**.** *Randomized experiments are sometimes used to learn about social interaction effects (e.g. Duflo and Saez 2003, Hirano and Hahn 2010). Let Z_1 and Z_2 be treatments applied to persons 1 and 2, respectively. Assuming the coefficients represent time-invariant structural parameters, random assignment of treatments ensures that the independence assumption A3 holds. If the treatment variable also satisfies the exclusion restriction, and a support condition (such as A4 or A4′ below), then I show one can identify the distribution of social interaction effects with experimental data.*

For example, suppose we are interested in learning the effect of student 1's GPA on their best friend student 2's GPA. Let our treatment Z_1 be the dollar value of a cash transfer paid to person 1 if they achieve a prespecified GPA cutoff. Likewise for Z_2. By incentivizing effort, larger values of the cash transfer Z_1 may induce person 1 to get a higher GPA. By randomly assigning different dollar values to the students, we can ensure that A3 holds.

[3] One exception is Heckman and Vytlacil (1998), who allow a specific kind of correlated random coefficient, although their goal is identification of the coefficients' means, not their distributions. Heckman, Schmierer, and Urzua (2010) construct tests of the independence assumption, building on earlier work by Heckman and Vytlacil (2007). Several papers, such as Graham and Powell (2012) and Arellano and Bonhomme (2012), relax independence by considering panel data models.

[4] For example, Matzkin (2003), Imbens and Newey (2009), Chernozhukov and Hansen (2005), Matzkin (2008, 2012) and Berry and Haile (2014), among others.

Assumption A4 (Instruments have continuous variation). $\text{supp}(Z_1 \mid X = x, Z_2 = z_2)$ contains an open ball in $\mathbb{R}$, for at least some $z_2 \in \text{supp}(Z_2 \mid X = x)$, for each $x \in \text{supp}(X)$. Likewise, $\text{supp}(Z_2 \mid X = x, Z_1 = z_1)$ contains an open ball in $\mathbb{R}$, for at least some $z_1 \in \text{supp}(Z_1 \mid X = x)$, for each $x \in \text{supp}(X)$.

This assumption requires that, conditional on one of the instruments and the other covariates, there must always be some region where we can vary the other instrument continuously. For example, it holds if $\text{supp}(Z \mid X = x)$ contains an open ball in $\mathbb{R}^2$, for each $x \in \text{supp}(X)$. It holds if $\text{supp}(Z \mid X) = \text{supp}(Z_1 \mid X) \times \text{supp}(Z_2 \mid X)$, where $\text{supp}(Z_1 \mid X)$ and $\text{supp}(Z_2 \mid X)$ are non-degenerate intervals. A4 also allows mixed continuous-discrete distributions, and it also allows the support of Z_1 to depend on the realization of Z_2, and vice versa. Moreover, as in the discussion following assumption A2, if we are only interested in one equation, then we do not need an instrument for the other equation. For example, suppose we only have the instrument Z_1 but not Z_2. Then we only need $\text{supp}(Z_1 \mid X = x)$ to contain an open ball in $\mathbb{R}$ to identify the distribution of $\gamma_2 \mid X$. For simplicity, the results here are stated under the assumption that we have an instrument for both equations.

Assumption A5 (Moment determinacy). Let π_i denote the vector of reduced form coefficients from equation $i = 1, 2$. These are defined shortly below.

1. Conditional on $X = x$, the absolute moments of the reduced form coefficients (π_1, π_2),

$$\int |p_1|^{\alpha_1} \cdots |p_6|^{\alpha_6} \, dF_{\pi_1, \pi_2 | X}(p \mid x), \qquad \alpha \in \mathbb{N}^6,$$

   are finite, for each $x \in \text{supp}(X)$. $\mathbb{N}$ denotes the natural numbers.

2. The distribution of $(\pi_1, \pi_2) \mid X = x$ is uniquely determined by its moments, for each $x \in \text{supp}(X)$.

As theorem 2 below shows, the necessity result from lemma 2 in single equation models carries over to simultaneous equations models: the tail conditions A5 are necessary if we wish to obtain point identification. A5 places restrictions directly on the reduced form coefficients π_i, rather than on the structural variables (B, D, U, Γ). A6 below provides sufficient conditions for A5, stated in terms of the structural variables directly. A5.1 implies that the reduced form mean regressions exist. It restricts the probability of nearly parallel lines (see section 2.2). Assumptions like A5.2 have been used in several papers to achieve identification, since it reduces the problem of identifying an entire distribution to that of identifying just its moments. For example, Fox, Kim, Ryan, and Bajari (2012) use it to identify a random coefficients logit model, and Ponomareva (2010) uses it to identify a quantile regression panel data model. A5.2 is a thin tail restriction on $\pi_i \mid X$; for example, any compactly supported distribution is uniquely determined by its moments, as well as any distribution whose moment generating function exists, like the normal distribution.

A simple sufficient condition for A5 is that the outcomes (Y_1, Y_2) have bounded support. This is often the case in practice, such as in the empirical application in section 5 where outcomes are GPAs. Alternatively, the sufficient conditions given in A6 below allow for outcomes to have full support, so long as their tails are thin enough (e.g., normally distributed). Consequently, it's only in applications where we expect outcomes to have fat tails where we might expect A5 to fail.

Theorem 2. Under A1, A2, A3, A4, and A5, the conditional distributions $\gamma_1 \mid X = x$ and $\gamma_2 \mid X = x$ are identified for each $x \in \text{supp}(X)$. Moreover, if $\text{supp}(Z \mid X = x)$ is bounded, then A5 is necessary for point identification of these marginal distributions.

The full proof is in appendix A. The main idea is as follows. The reduced form system (3) is

$$Y_1 = \frac{U_1 + \gamma_1 U_2 + (\delta_1 + \gamma_1 \delta_2)' x}{1 - \gamma_1 \gamma_2} + \frac{\beta_1}{1 - \gamma_1 \gamma_2} Z_1 + \frac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2} Z_2 \equiv \pi_{11} + \pi_{12} Z_1 + \pi_{13} Z_2$$
$$Y_2 = \frac{U_2 + \gamma_2 U_1 + (\delta_2 + \gamma_2 \delta_1)' x}{1 - \gamma_1 \gamma_2} + \frac{\gamma_2 \beta_1}{1 - \gamma_1 \gamma_2} Z_1 + \frac{\beta_2}{1 - \gamma_1 \gamma_2} Z_2 \equiv \pi_{21} + \pi_{22} Z_1 + \pi_{23} Z_2.$$

For $(t_1, t_2) \in \mathbb{R}^2$, we have

$$t_1 Y_1 + t_2 Y_2 = (t_1 \pi_{11} + t_2 \pi_{21}) + (t_1 \pi_{12} + t_2 \pi_{22}) Z_1 + (t_1 \pi_{13} + t_2 \pi_{23}) Z_2.$$

Condition on $Z_1 = z_1$. Then by applying lemma 2 on identification of random coefficients in single equation models, we can identify the joint distribution of

$$([t_1 \pi_{11} + t_2 \pi_{21}] + [t_1 \pi_{12} + t_2 \pi_{22}] z_1, \quad t_1 \pi_{13} + t_2 \pi_{23})$$

for any $(t_1, t_2) \in \mathbb{R}^2$. This lets us learn the joint distribution of, for example,

$$(\pi_{13}, \pi_{23}) = \left(\frac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2}, \frac{\beta_2}{1 - \gamma_1 \gamma_2} \right) \tag{5}$$

and from this we have $\gamma_1 = \pi_{13}/\pi_{23}$. Similarly, if we first condition on $Z_2 = z_2$ instead of $Z_1 = z_1$ then we can identify the joint distribution of (π_{12}, π_{22}) and thus the distribution of γ_2. This proof strategy is analogous to a standard approach for constant coefficient simultaneous equations models, in which case π_{13} and π_{23} are constants whose ratio equals the constant γ_1. The necessity of A5 follows since the simultaneous equations model nests the single equation model, and A5 is necessary for identification of the marginal distributions of regressor coefficients by lemma 2.

Recall that in the proof of lemma 2 we do not need to assume that the distribution of random coefficients is uniquely determined by its moments (assumption (4) in lemma 2) if we only wish to identify moments of the distribution of coefficients. So, in the simultaneous equations model, if we eliminate assumption A5.2, then we can still identify all moments of π_1 and π_2. Unfortunately, these reduced form moments do not necessarily identify the structural moments $\mathbb{E}(\gamma_1 \mid X)$ and $\mathbb{E}(\gamma_2 \mid X)$, assuming these structural moments exist.

The only restrictions on the joint distribution of unobservables (B, D, U, Γ) used in theorem 2 are the unique solution assumption A1 and the moment determinacy condition A5. Unlike earlier work such as Kelejian (1974), these conditions do not require the unobservables to be independent of each other. Allowing for dependence is important in many applications, such as the following.

Example (Social interactions between pairs of people, cont'd)**.** *Suppose we examine social interactions between best friend pairs. Friendships may form because a pair of students have similar observed and unobserved variables. Consequently we expect that* $(\beta_1, \delta_1, \gamma_1, U_1)$ *and* $(\beta_2, \delta_2, \gamma_2, U_2)$ *are not independent. These are called correlated effects in the social interactions literature. Such dependence is fully allowed here when identifying the distributions of social interaction effects* γ_1 *and* γ_2*. Furthermore, the covariates* X*, which may contain variables like person 1's gender and person 2's gender, can be arbitrarily related to the unobservables.*

Recall the following definition: Suppose a random variable V satisfies

$$\mathbb{P}(|V| > t) \leq C \exp(-ct^p)$$

for some constants $C, c > 0$ that depend on V but not t. If $p = 1$ we say V has *subexponential tails* while if $p = 2$ we say V has *sub-Gaussian tails*. Then a sufficient condition for A5, in terms of the structural parameters, is the following.

Assumption A6 (Restrictions on structural unobservables).

1. $\mathbb{P}(|1 - \gamma_1\gamma_2| \geq \tau \mid X) = 1$ for some $\tau > 0$. That is, $1 - \gamma_1\gamma_2$ is bounded away from zero. Equivalently, $\gamma_1\gamma_2$ is bounded away from 1.

2. Conditional on X, the distributions of β_1, β_2, $\beta_1\gamma_2$, $\beta_2\gamma_1$,$(U_1, \delta_1, \gamma_1 U_2, \gamma_1\delta_2)$,$(U_2, \delta_2, \gamma_2 U_1, \gamma_2\delta_1)$ have subexponential tails.

Proposition 1. A6 implies A5.

As noted earlier, A5 is necessary for point identification. Hence that is the weakest possible set of assumptions on the distribution of structural unobservables we can make while still achieving point identification. Assumption A6 strengthens A5 slightly in order to obtain more interpretable conditions. The main difference is that while A5 allows $1 - \gamma_1\gamma_2$ to be arbitrarily close to zero, so long as it has sufficiently little mass near zero, assumption A6.1 rules this out. A6.1 holds if γ_1 and γ_2 are always known to have opposite signs, as in the supply and demand example, or if the magnitude of both γ_1 and γ_2 is bounded above by some $\tau < 1$ (see proposition 5 in appendix A). The latter assumption may be reasonable in social interactions applications, where a positive social interaction coefficient of 1 or greater would be substantively quite large and perhaps unlikely to be true; also see the discussion of stability below.

A6.2 requires certain structural unobservables and certain cross-products of these random variables to have thin enough tails. This tail condition accommodates most well known distributions,

such as the normal distribution, as well as any compactly supported distribution. Also, as used in the proof of proposition 1, a random variable having subexponential tails is equivalent to that variable's moment generating function existing in a neighborhood of zero. This is an equivalent way to view the tail restriction in A6.2.

A6.2 is stated in terms of products of structural unobservables. The following result gives two different sets of sufficient conditions for assumption A6.2. These conditions do not involve products and hence are even simpler to interpret, although they are not necessary for point identification.

Proposition 2. Assume either

> (A6.2′) Conditional on X, the marginal distributions of all the structural random variables γ_1, γ_2, β_1, β_2, U_1, U_2, $\delta_{11}, \ldots, \delta_{1K}$, $\delta_{21}, \ldots, \delta_{2K}$ have sub-Gaussian tails.

or

> (A6.2″) Conditional on X, γ_1 and γ_2 have compact support and, conditional on X, the distributions of β_1, β_2, and $(U_1, U_2, \delta_1, \delta_2)$ have subexponential tails.

Then A6.2 holds.

A6.2 requires subexponential tails for products of random variables. This proposition therefore shows the tradeoff between the relative tails of the two random variables being multiplied. Compact support for the endogenous variable random coefficients allows the remaining unobservables to have merely subexponential tails, while if we allow the endogenous variable random coefficients to have full support and sub-Gaussian tails, we must restrict the the remaining unobservables to have thinner than subexponential tails.

A6.1 rules out distributions of (γ_1, γ_2) with support such that $\gamma_1\gamma_2$ is arbitrarily close to one. In particular, this rules out distributions with full support, like the bivariate normal (although it allows for Gaussian tails). While the normal distribution is often used in applied work for random coefficients on exogenous variables, it has perhaps unappealing implications as a distribution of random coefficients on endogenous variables in models with simultaneity. First, it can easily lead to distributions of $1/(1-\gamma_1\gamma_2)$ which have no moments, and hence outcome variables which have no moments. For example, suppose γ_2 is a constant coefficient and $\gamma_1 \sim \mathcal{N}(\mu, \sigma^2)$. Then $1/(1-\gamma_1\gamma_2) \sim 1/\mathcal{N}(1-\gamma_2\mu, \gamma_2^2\sigma^2)$, which does not have a mean (see example (a) on page 40 of Robert 1991). Consequently, normally distributed coefficients (γ_1, γ_2) are unlikely to be consistent with the data if our outcomes have at least one moment. Moreover, if $1/(1-\gamma_1\gamma_2)$ has no moments, then A5 may fail. For example, if β_1 is constant then $\pi_{12} = \beta_1/(1-\gamma_1\gamma_2)$ would have no moments. This then implies that the marginal distributions of endogenous variable coefficients are not point identified, since A5 is necessary for point identification.

Second, one may find it reasonable to assume that the equilibrium in system (1) is stable, in the sense that if we perturb the equilibrium, the system returns back to equilibrium instead of

than diverging to infinity. Formally, consider a single realization of the unobservables (Γ, B, D, U). Although there are many ways to model disequilibrium dynamics, consider the simple dynamic process

$$Y_t = \Gamma Y_{t-1} + BZ + DX + U$$

for each time period $t = 1, 2, \ldots$, where Y_0 is some initial value (or the point which we perturb to). Let

$$Y = (I - \Gamma)^{-1} BZ + (I - \Gamma)^{-1} DX + (I - \Gamma)^{-1} U$$

denote the equilibrium (or steady state) value of outcomes. Say this equilibrium is *globally stable* if for any $Y_0 \in \mathbb{R}^2$, $Y_t \to Y$ as $t \to \infty$. The equilibrium is globally stable if and only if $|\gamma_1 \gamma_2| < 1$ (see appendix A). For example, a sufficient condition is that $|\gamma_1| < 1$ and $|\gamma_2| < 1$. As discussed above, this is perhaps reasonable in social interactions applications. See, for example, Bramoullé, Kranton, and D'Amours (2014) and Bramoullé and Kranton (2015) for further discussion of stability in this context. Any distribution of (γ_1, γ_2) with full support, such as the normal distribution, implies that a positive proportion of systems are globally unstable.

Finally, the question of which specific distributions of random coefficients on the endogenous variable are reasonable to allow is related to problems encountered when choosing priors for Bayesian analysis of the constant coefficient simultaneous equations model. Kleibergen and van Dijk (1994) showed that a diffuse prior on the structural coefficients leads to nonintegrability in the posterior. Chao and Phillips (1998, section 6) give more details and propose using a prior that avoids this thick tail problem.

Theorem 2 allows for instruments with bounded support. If our instruments have unbounded support, then we no longer need the moment determinacy conditions A5.

Assumption A4′ (Full, rectangular support instruments). $\text{supp}(Z \mid X) = \mathbb{R}^2$.

Theorem 3. Under A1, A2, A3, and A4′, the conditional distributions $\gamma_1 \mid X = x$ and $\gamma_2 \mid X = x$ are identified for each $x \in \text{supp}(X)$.

The proof is essentially identical to that of theorem 2. The only difference is that in the first step we apply a different identification result for the single-equation random coefficient model, namely, lemma 1 rather than lemma 2.[5]

The following result shows that the full joint distribution of structural unobservables is not point identified, even with full support instruments and assuming the moment determinacy conditions.

Theorem 4. Under A1, A2, A3, A4′, and A5, the joint distribution of (B, D, Γ, U) is not point identified. If we further assumed that U_1, U_2, and D are degenerate on zero, then the joint distribution of $(B, \Gamma) = (\beta_1, \beta_2, \gamma_1, \gamma_2)$ is still not point identified.

[5] A referee pointed out the following alternative proof, which only requires the conditional support of the instruments to be unbounded, but not necessarily all of $\mathbb{R}$: Y_1/Y_2 can be written as $(\gamma_1 + V_1/Z_2)/(1 + V_2/Z_2)$ for some random variables (V_1, V_2) and hence $\mathbb{P}(Y_1/Y_2 \leq t \mid X = x, Z_1 = z_1, Z_2 = z_2) \to \mathbb{P}(\gamma_1 \leq t \mid X = x)$ as $z_2 \to \pm\infty$. Likewise for the distribution of $\gamma_2 \mid X$.

Proof of theorem 4. The system of reduced form equations (3) is a SUR model whose regressors are common across equations. Hence corollary 1 to theorem 1 on SUR models implies that the joint distribution of reduced form coefficients is not point identified. There is a one-to-one mapping between the reduced form coefficients and the structural coefficients (B, D, Γ, U). Consequently, the joint distribution of structural coefficients is not point identified.

The second result follows because U_1, U_2, and D degenerate on zero does not change the fact that there are common regressors across equations, and so the joint distribution of the four remaining reduced form coefficients is still not point identified. □

Theorems 2 and 3 show that, despite this nonidentification result, we are able to point identify the marginal distributions of endogenous variable random coefficients.

Next I consider identification of the joint distribution of endogenous variable coefficients. In some cases the empirical setting naturally provides additional restrictions on the joint distribution of γ_1 and γ_2, as in the following example.

Example (Social interactions between pairs of people, cont'd)**.** *Assuming the unobservables represent time-invariant structural parameters, independence between $(\beta_1, \delta_1, \gamma_1, U_1)$ and $(\beta_2, \delta_2, \gamma_2, U_2)$ holds when people are randomly paired, as in laboratory experiments (e.g. Falk and Ichino 2006) or natural experiments (e.g. Sacerdote 2001). In particular, there is no matching based on the endogenous social interaction effect; γ_1 and γ_2 are independent.*

In other cases, however, we might expect γ_1 and γ_2 to be correlated. In this case, identification of the joint distribution of (γ_1, γ_2) would, for example, allow us to learn whether assortative matching between friends occurred along the dimension of social susceptibility. The following result shows that when the instrument coefficients are constant, we are able to identify this joint distribution.

Proposition 3. Assume the conditions of theorem 2 hold. Suppose further that (i) β_1 and β_2 are constant coefficients, (ii) $\mathbb{P}(\gamma_1\gamma_2 < 1 \mid X) \neq \mathbb{P}(\gamma_1\gamma_2 > 1 \mid X)$, and (iii) A4 is strengthened to $\text{supp}(Z_1, Z_2 \mid X)$ contains an open ball in $\mathbb{R}^2$. Then for each $x \in \text{supp}(X)$, the joint distribution of $(\gamma_1, \gamma_2) \mid X = x$ is identified. If we also assume (iv) $\mathbb{E}[1/(1 - \gamma_1\gamma_2)]$ exists and is nonzero, then β_1 and β_2 are identified.

The idea here is that assuming β_1 and β_2 are constant reduces the dimension of unobserved heterogeneity in the reduced form coefficients on the instruments,

$$(\pi_{12}, \pi_{22}, \pi_{13}, \pi_{23}) = \left(\frac{\beta_1}{1 - \gamma_1\gamma_2}, \frac{\gamma_1\beta_2}{1 - \gamma_1\gamma_2}, \frac{\gamma_2\beta_1}{1 - \gamma_1\gamma_2}, \frac{\beta_2}{1 - \gamma_1\gamma_2} \right),$$

from 4 to 2. Moreover, the distribution of 'own' coefficients (the effect of Z_i on Y_i) are just scaled versions of each other:

$$\pi_{23} = \pi_{12} \frac{\beta_2}{\beta_1}.$$

With these observations the result follows from modifying the proof strategy for theorem 2.

Under the assumption that β_1 and β_2 are constant, the relevance assumption A2 simply states that β_1 and β_2 are nonzero. Assumption (ii) here restricts the amount of symmetry in the distribution of $\gamma_1\gamma_2$—it cannot have equal mass both below and above 1. If the distribution of $1-\gamma_1\gamma_2$ is continuous and has a strictly increasing cdf, then assumption (ii) is equivalent to the assumption that the median of $1-\gamma_1\gamma_2$ cannot be 0. Assumption (ii) is only used to identify the sign of β_1/β_2. If this sign is known a priori then assumption (ii) is not needed. For example, if it is known that $\beta_1 = \beta_2$ (for example, as in the best friend pairs example, since the labels of friend 1 and friend 2 do not matter), then β_1/β_2 is known to be positive. Assumption (iii) is used here because it lets us identify the joint distribution of the linear combinations of reduced form coefficients on different instruments, which we use to recover the joint distribution of (γ_1, γ_2).

Assumption (iv) is only used for identifying the sign of β_1 and the sign of β_2; it is not used to identify the joint distribution of (γ_1, γ_2). Assumption (iv) is also a restriction on the symmetry of the distribution of $\gamma_1\gamma_2$. Assumption (iv) holds in some common cases, like with supply and demand, where we know that $\gamma_1\gamma_2 \leq 0$, since supply slopes up and demand slopes down, and hence $1-\gamma_1\gamma_2 > 0$ so the mean of the inverse must be strictly positive. See proposition 5 in the appendix for more discussion of sufficient conditions for assumptions (ii) and (iv).

I conclude this section with a result on triangular systems and a remark about additive separability and linearity. The following result uses the proof of either theorem 2 or 3 to examine triangular systems, a case of particular relevance for the literature on heterogeneous treatment effects.

Proposition 4. Consider model (1) with β_1 and γ_2 degenerate on zero:

$$\begin{aligned} Y_1 &= \gamma_1 Y_2 + \delta_1' X + U_1 \\ Y_2 &= \beta_2 Z_2 + \delta_2' X + U_2. \end{aligned} \tag{6}$$

Assume

1. (Relevance) $\mathbb{P}(\beta_2 = 0 \mid X) = 0$
2. (Independence) $Z_2 \perp\!\!\!\perp (\gamma_1, \beta_2, \delta_1, \delta_2, U_1, U_2) \mid X$

and either

3′. (Full support instruments) $\operatorname{supp}(Z_2 \mid X) = \mathbb{R}$

or

3. (Instruments have continuous variation) $\operatorname{supp}(Z_2 \mid X)$ contains an open ball in $\mathbb{R}$
4. (Moment determinacy) The distribution of

$$(U_1 + \gamma_1 U_2 + (\delta_1 + \gamma_1\delta_2)'x, U_2 + \delta_2' x, \beta_1, \beta_2, \gamma_1\beta_2) \mid X = x$$

has finite absolute moments and is uniquely determined by its moments, for each $x \in \text{supp}(X)$. Then the joint distribution of $(\gamma_1, \beta_2) \mid X$ is identified. Moreover, if $\text{supp}(Z_2 \mid X)$ is bounded, then the moment determinacy assumption is necessary for identification of the marginal distribution of $\gamma_1 \mid X$ and the marginal distribution of $\beta_2 \mid X$.

For example, suppose Y_1 is log-wage and Y_2 is education. While the 2SLS estimator of γ_1 in the triangular model (6) converges to a weighted average effect parameter, this proposition provides conditions for identifying the distribution of treatment effects, $\gamma_1 \mid X$. The assumption that β_1 is degenerate on zero just means that no instrument Z_1 for the first stage equation is required for identification, as usual with triangular models; any variables Z_1 excluded from the first stage equation may be included in X by making appropriate zero restrictions on δ_2. Proposition 4 makes no restrictions on the dependence structure of the unobservables $(U_1, U_2, \gamma_1, \beta_2, \delta_1, \delta_2)$, which allows (6) to be a correlated random coefficient model. For example, education level Y_2 may be chosen based on one's individual-specific returns to education γ_1, which implies that (β_2, δ_2, U_2) and γ_1 would not be independent. Sufficient conditions for the moment determinancy assumption can be obtained by applying propositions 1 and 2. For example, moment determinacy in the triangular model holds if all the structural unobservables have sub-Gaussian tails. Hoderlein et al. (2010, page 818) also discuss identification of a triangular model like (6), but they assume β_2 is constant.

Remark 2 (The role of additive separability and linearity)**.** In both systems (1) and (6), the exogenous covariates X are allowed to affect outcomes directly via an additive term and indirectly via the random coefficients. Without further restrictions on the effect of X, the inclusion of δ_1 and δ_2 is redundant. We could instead rewrite the system as

$$\begin{aligned} Y_1 &= \gamma_1(X)Y_2 + \beta_1(X)Z_1 + V_1(X) \\ Y_2 &= \gamma_2(X)Y_1 + \beta_2(X)Z_2 + V_2(X), \end{aligned}$$

where $\gamma_i(\cdot)$, $\beta_i(\cdot)$, and $V_i(\cdot)$ are arbitrary random functions of X, $i = 1, 2$. This formulation emphasizes that the key functional form assumption is that the endogenous variables and the instruments to affect outcomes linearly. Nonetheless, system (1) is more traditional, and is also helpful when proceeding to estimation where we make assumptions on the effect of X for dimension reduction. □

3.4 Many equations: Social interactions models

In this section I discuss several extensions to systems of more than two equations. For simplicity I omit covariates X throughout this section; all assumptions and results can be considered as conditional on X. A general linear system of N simultaneous equations can be written as

$$Y_i = \sum_{i=1}^{N} \gamma_{ij} Y_j + \beta_i Z_i + U_i \tag{7}$$

for $i = 1, \ldots, N$, and $\gamma_{ii} = 0$. As before, the β_i and U_i are unobserved random variables. In this case, there are $N(N-1) = O(N^2)$ random coefficients on the endogenous variables. Without further assumptions, it is generally not possible to identify the entire joint distribution of all these coefficients. Consequently, in this section I consider restrictions on the set of coefficients $\{\gamma_{ij}\}$ which yield point identification of distributions of coefficients. There are many possible restrictions one could consider, depending on the empirical context. I will focus on applications to social interactions models. I begin with a random coefficients generalization of the most widely used social interactions model, the linear-in-means model (Manski 1993). I then consider a generalization which incorporates observed network data. In both cases I give sufficient conditions for point identification of the marginal distributions of endogenous variable coefficients.

3.4.1 The linear-in-means model with heterogeneous social interaction effects

The classical linear-in-means model assumes that each person i's outcome is a linear function of the average of all other persons in their reference group:

$$Y_i = \theta \frac{1}{N-1} \sum_{j \neq i} Y_j + \beta_i Z_i + U_i.$$

θ is called the *endogenous social interaction* parameter. See Blume, Brock, Durlauf, and Ioannides (2011) for a survey and many further references. Typically these models assume that this parameter is constant and common to all units. In this section, I consider the case where θ is a random coefficient, which allows for heterogeneous social interaction effects. Different people may be influenced by the mean of their peer group differently. Specifically, I consider the model

$$Y_i = \gamma_i \frac{1}{N-1} \sum_{j \neq i} Y_j + \beta_i Z_i + U_i \tag{2}$$

mentioned on page 3, where γ_i is a random coefficient. This can be obtained from equation (7) by assuming that, for each i, $\{\gamma_{ij} : j = 1, \ldots, N, j \neq i\}$ are all equal to a single random variable γ_i. Thus the number of unknown random coefficients is reduced from $O(N^2)$ to $O(N)$, which turns out to be sufficient to achieve point identification of the marginal distribution of each γ_i. Notice that in this social interactions example, we typically think that the labels of people in the group are arbitrary, and hence expect the marginal distributions of all the γ_i should be identical. This assumption is not needed for the identification argument, however.

I have omitted exogenous social interactions effects from the model. These occur when person j's covariates X_j affect the outcome Y_i of person i. These may be included without affecting the main results below; indeed, each person j's covariates X_j may enter person i's outcome equation with its own random coefficients δ_{ij}. The key assumption, however, is that I do *not* allow exogenous social interaction effects of the instruments $Z = (Z_1, \ldots, Z_N)$. That is, there is at least one covariate that affects i's outcome but no one else's. This assumption is similar to $\gamma = 0$ in Manski's (1993)

proposition 2; also see Brock and Durlauf (2001, page 3324) and Evans, Oates, and Schwab (1992).

As earlier, let $B = \text{diag}(\beta_1, \ldots, \beta_N)$, $U = (U_1, \ldots, U_N)$, and Γ denote the matrix of random coefficients on the endogenous variables.

Theorem 5. Consider the linear-in-means model (2). Assume

1. (Support of endogenous effects) There is a $\tau \in (0, 1)$ such that $\mathbb{P}(|\gamma_i| \leq \tau) = 1$ for all $i = 1, \ldots, N$.
2. (Relevance) $\mathbb{P}(\beta_i = 0) = 0$ for all $i = 1, \ldots, N$.
3. (Independence) $Z \perp\!\!\!\perp (B, U, \Gamma)$.
4. (Instruments have continuous variation) $\text{supp}(Z_i \mid Z_{-i} = z_{-i})$ contains an open ball in $\mathbb{R}$ for at least some $z_{-i} \in \text{supp}(Z_{-i})$, for all $i = 1, \ldots, N$, where $Z_{-i} = \{Z_k : k \neq i\}$.

Then the joint distribution of any subset of $N - 1$ elements of $\{\gamma_1, \ldots, \gamma_N\}$, is point identified. In particular, the marginal distribution of γ_i is point identified, for each $i = 1, \ldots, N$.

Assumptions (2)–(4) are as in the two equation case. The main new assumption here is (1), which restricts the support of the random coefficients γ_i to be in $(-\tau, \tau) \subsetneq (-1, 1)$. Previous research often assumes a common, constant endogenous social interaction coefficient θ such that $|\theta| < 1$ (e.g., Case 1991, Bramoullé, Djebbari, and Fortin 2009, and Blume, Brock, Durlauf, and Jayaraman 2015). Hence assumption (1) is a strict generalization of this previous assumption. The random coefficients linear-in-means model here has similar benefits as in the two equation case. It does not require all people to be positively affected by their peers. Likewise, it does not require all people to be negatively affected by their peers. Some people may have positive effects while others may have negative effects. Moreover, some people may be strongly affected by their peers (large γ_i) while others may be only moderately affected by their peers (small γ_i).

The interpretation of assumption (1) is similar to that discussed in proposition 5 (in the appendix) in the two equation case: Variation in the mean outcomes of i's peers will never change i's outcome Y_i by larger than the magnitude change in mean peer outcomes. For example, if the mean GPA in my peer group increases by 1 point, my GPA will not increase by more than 1 point, and it will not decrease by more than 1 point.

Assumption (1) implies that the reduced form system exists with probability 1. It also ensures that the unique equilibrium is stable. Finally, it ensures that the moments of the distribution of reduced form coefficients all exist and uniquely determine that distribution, and it also ensures that certain random variables are bounded away from zero as used in the proof.

The full proof of theorem 5 is in the appendix. To see the main idea, consider the following

three equation system:

$$
\begin{aligned}
Y_1 &= \gamma_1 \left(\frac{Y_2 + Y_3}{2} \right) + \beta_1 Z_1 + U_1 \\
Y_2 &= \gamma_2 \left(\frac{Y_1 + Y_3}{2} \right) + \beta_2 Z_2 + U_2 \\
Y_3 &= \gamma_3 \left(\frac{Y_1 + Y_2}{2} \right) + \beta_3 Z_3 + U_3.
\end{aligned}
$$

The reduced form is

$$
\begin{aligned}
Y_1 &= \det(\widetilde{\Gamma})^{-1} \left[\left(1 - \gamma_2\gamma_3 \frac{1}{4} \right) \beta_1 Z_1 + \gamma_1 \left(\frac{1}{2} + \gamma_3 \frac{1}{4} \right) \beta_2 Z_2 + \gamma_1 \left(\frac{1}{2} + \gamma_2 \frac{1}{4} \right) \beta_3 Z_3 + \cdots \right] \\
Y_2 &= \det(\widetilde{\Gamma})^{-1} \left[\gamma_2 \left(\frac{1}{2} + \gamma_3 \frac{1}{4} \right) \beta_1 Z_1 + \left(1 - \gamma_1\gamma_3 \frac{1}{4} \right) \beta_2 Z_2 + \gamma_2 \left(\frac{1}{2} + \gamma_1 \frac{1}{4} \right) \beta_3 Z_3 + \cdots \right] \\
Y_3 &= \det(\widetilde{\Gamma})^{-1} \left[\gamma_3 \left(\frac{1}{2} + \gamma_2 \frac{1}{4} \right) \beta_1 Z_1 + \gamma_3 \left(\frac{1}{2} + \gamma_1 \frac{1}{4} \right) \beta_2 Z_2 + \left(1 - \gamma_1\gamma_2 \frac{1}{4} \right) \beta_3 Z_3 + \cdots \right]
\end{aligned}
$$

where the omitted terms are random intercepts depending on (U_1, U_2, U_3) and $\widetilde{\Gamma} = I - \Gamma$. As in the two equation case, we can point identify the joint distribution of reduced form coefficients on Z_1:

$$
(\pi_{11}, \pi_{21}, \pi_{31}) \equiv \left(\left(1 - \gamma_2\gamma_3 \frac{1}{4} \right) \frac{\beta_1}{\det(\widetilde{\Gamma})}, \quad \gamma_2 \left(\frac{1}{2} + \gamma_3 \frac{1}{4} \right) \frac{\beta_1}{\det(\widetilde{\Gamma})}, \quad \gamma_3 \left(\frac{1}{2} + \gamma_2 \frac{1}{4} \right) \frac{\beta_1}{\det(\widetilde{\Gamma})} \right).
$$

By dividing the first coefficient into the other two, we point identify the joint distribution of

$$
\left(\frac{\pi_{21}}{\pi_{11}}, \frac{\pi_{31}}{\pi_{11}} \right) = \left(\frac{\gamma_2(1/2 + \gamma_3/4)}{1 - \gamma_2\gamma_3/4}, \frac{\gamma_3(1/2 + \gamma_2/4)}{1 - \gamma_2\gamma_3/4} \right).
$$

The point identified random variables $(\pi_{21}/\pi_{11}, \pi_{31}/\pi_{11})$ are a one-to-one mapping of the structural coefficients γ_2 and γ_3:

$$
\gamma_2 = \frac{2(\pi_{21}/\pi_{11})}{1 + (\pi_{31}/\pi_{11})} \qquad \text{and} \qquad \gamma_3 = \frac{2(\pi_{31}/\pi_{11})}{1 + (\pi_{21}/\pi_{11})}.
$$

Hence the joint distribution of (γ_2, γ_3) is point identified via a change of variables. The key observation here is that the reduced form coefficients on Z_1 depend on γ_1 only via the determinant term; γ_1 does not appear anywhere else. Consequently, when taking ratios both β_1 and γ_1 disappear from the subsequent expression. Intuitively, Z_1 is an instrument for the endogenous variable Y_1, and hence is used to identify the effects of Y_1 on the other outcome variables, Y_2 and Y_3; i.e., the random coefficients γ_2 and γ_3.

A similar argument can be applied to the reduced form coefficients on Z_2 to show that the joint distribution of (γ_1, γ_3) is point identified. Consequently, the marginal distributions of all random coefficients are point identified. The proof in the appendix shows that this argument extends to

systems of N equations.

3.4.2 The linear-in-means network model

A variation on the classical linear-in-means model discussed above takes means over an observed, person specific subset of people in the overall group, rather than including everyone in the mean (e.g., Bramoullé et al. 2009, Lee, Liu, and Lin 2010, and Blume et al. 2015). Specifically, suppose there are N people in a network. Then the linear-in-means network model specifies person i's outcome as

$$Y_i = \gamma_i \frac{1}{|\mathscr{N}(i)|} \sum_{j \in \mathscr{N}(i)} Y_j + \beta_i Z_i + U_i. \tag{8}$$

$\mathscr{N}(i)$ is an observed subset of the indices $\{j = 1, \ldots, N : j \neq i\}$, called the 'neighborhood' of person i. γ_i is a random coefficient that represents the effect of the average outcome within person i's neighborhood on Y_i. Let $N_i = |\mathscr{N}(i)|$ denote the number of people who influence person i. Let $1_{ij} = \mathbb{1}[j \in \mathscr{N}(i)]$ denote the indicator of whether j is person i's neighborhood. Let A denote the matrix whose ij-th element is 1_{ij}. A is called the *adjacency matrix*. Assume that A is an observable random matrix.

The following result generalizes theorem 5.

Theorem 6. Consider model (8). Assume

1. (Support of endogenous effects) There is a $\tau \in (0, 1)$ such that $\mathbb{P}(|\gamma_i| \leq \tau \mid A) = 1$ for all $i = 1, \ldots, N$.
2. (Relevance) $\mathbb{P}(\beta_i = 0 \mid A) = 0$ for all $i = 1, \ldots, N$.
3. (Independence) $Z \perp\!\!\!\perp (B, U, \Gamma) \mid A$.
4. (Instruments have continuous variation) $\text{supp}(Z \mid A)$ contains an open ball in $\mathbb{R}^N$.
5. (Everyone has a friend) $\mathbb{P}(N_i \geq 1) = 1$ for all $i = 1, \ldots, N$.

Then the marginal distribution of $\gamma_i \mid A$ is point identified for each $i = 1, \ldots, N$.

The assumptions here are similar to those of theorem 5. The main difference is that now we are conditioning on the adjacency matrix A. Hence, for the identification analysis, it is not necessary for the links to be formed independently of the unobservables (B, U, Γ), so long as the instruments are statistically independent of the unobservables conditional on A. Moreover, I also assume that everyone is influenced by at least one person simply to rule out the trivial cases where the distribution of $\gamma_i \mid A$ is not identified because we are looking only at networks where Y_i is not influenced by anyone, in which case γ_i would not enter the outcome equation (8).

The proof is similar to that of theorem 5. Consider the $N = 3$ case. The vector of coefficients on Z_1 is

$$(\pi_{11}, \pi_{21}, \pi_{31}) = \frac{\beta_1}{\det(\widetilde{\Gamma})} \left(1 - \frac{\gamma_2}{N_2}\frac{\gamma_3}{N_3} 1_{23} 1_{32}, \quad \frac{\gamma_2}{N_2}\left(1_{21} + 1_{23} 1_{31} \frac{\gamma_3}{N_3}\right), \quad \frac{\gamma_3}{N_3}\left(1_{31} + 1_{32} 1_{21} \frac{\gamma_2}{N_2}\right)\right).$$

Dividing the first component into the second and third components cancels out the determinant and β_1 terms, and yields a system of two reduced form random variables in the two structural random variables γ_2 and γ_3. This system can be solved for to get:

$$\gamma_2 = \frac{N_2(\pi_{21}/\pi_{11})}{1_{21} + 1_{23}(\pi_{31}/\pi_{11})} \qquad \text{and} \qquad \gamma_3 = \frac{N_3(\pi_{31}/\pi_{11})}{1_{31} + 1_{32}(\pi_{21}/\pi_{11})}.$$

The main difference with theorem 5 is that the matrix Γ of random coefficients has some zero terms, where $1_{ij} = 0$, and we let the denominator of the weights be $|\mathscr{N}(i)|$ instead of $N - 1$. Moreover, note that in order to get the distribution of γ_2 and γ_3 from these expressions, we need the reduced form effects of person 1 on 2, π_{21}, and of person 1 on 3, π_{31}, to be nondegenerate. This is guaranteed in the linear-in-means model, but not in this directed network model. For example, consider the network where persons 2 and 3 are influenced by each other, but not by person 1. And person 1 is influenced by 2 and 3. In this case, $\pi_{21} \equiv \pi_{31} \equiv 0$ since $1_{21} = 1_{31} = 0$. Consequently, the above expressions would not identify the distribution of γ_2 and γ_3. This is intuitive because above we are looking at the effects of Z_1 on (Y_1, Y_2, Y_3), and yet we know that person 1 does not influence 2 and 3. Instead, because we know that 2 affects 3, we can look at the effect of Z_2 on (Y_1, Y_2, Y_3) instead. In this case, we know that both π_{22} and π_{32} are nondegenerate, and similar derivations to those above show that we can identify the distribution of γ_2.

Finally, consider the question of learning the joint distribution of γ_j and γ_k, as in theorem 5. If we further assume that there is a person i who has at least an indirect effect on both j and k, then we can identify the joint distribution of (γ_j, γ_k). This can be seen in the above three person example by letting $j = 2$, $k = 3$, and $i = 1$. As before, this argument can be used to get the joint distribution of at most $N - 1$ of the endogenous variable coefficients.

4 Estimation

In this section I consider estimation of the marginal distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$ in system (1), under the identification assumptions of section 3. While I describe the estimator for two equation systems, the approach can be generalized to the many equation setting. I first describe the estimator. I then examine the estimator's finite sample performance with several Monte Carlo simulations. I end by discussing bandwidth selection in practice.

4.1 Nonparametric estimation

In this section I describe a constructive, nonparametric kernel-based estimator which is a sample analog to the identification arguments. For simplicity I omit covariates X. It's straightforward to include them in step 1 below, and I also discuss a single-index approach to including covariates below. I focus on estimating the pdf of γ_2. The approach for γ_1 is analogous.

Recall that the reduced form of system (1) is

$$Y_1 = \frac{U_1 + \gamma_1 U_2}{1 - \gamma_1\gamma_2} + \frac{\beta_1}{1 - \gamma_1\gamma_2} Z_1 + \frac{\gamma_1\beta_2}{1 - \gamma_1\gamma_2} Z_2 \equiv \pi_{11} + \pi_{12} Z_1 + \pi_{13} Z_2$$

$$Y_2 = \frac{\gamma_2 U_1 + U_2}{1 - \gamma_1\gamma_2} + \frac{\gamma_2\beta_1}{1 - \gamma_1\gamma_2} Z_1 + \frac{\beta_2}{1 - \gamma_1\gamma_2} Z_2 \equiv \pi_{21} + \pi_{22} Z_1 + \pi_{23} Z_2.$$

For $(t_1, t_2) \in \mathbb{R}^2$, we have

$$\begin{aligned} t_1 Y_1 + t_2 Y_2 &= (t_1\pi_{11} + t_2\pi_{21}) + (t_1\pi_{12} + t_2\pi_{22}) Z_1 + (t_1\pi_{13} + t_2\pi_{23}) Z_2 \\ &\equiv \Pi_1(t_1, t_2) + \Pi_2(t_1, t_2) Z_1 + \Pi_3(t_1, t_2) Z_2. \end{aligned}$$

Let

$$\Pi(t_1, t_2) \equiv \big(\Pi_1(t_1, t_2), \Pi_2(t_1, t_2), \Pi_3(t_1, t_2)\big)$$

denote the vector of random coefficients in this single equation model. The estimator has four steps, described as follows.

1. (Linear combination reduced form pdf) Apply an existing method (see the discussion below) to obtain $\widehat{f}_{\Pi(t_1,t_2)}$, an estimate of the pdf of $\Pi(t_1, t_2)$ of linear combinations of the reduced form coefficients. This is 3-dimensional in the two equation case with one instrument per equation and no covariates. In general, it is $1 + d_{Z_1} + d_{Z_2} + d_X$ dimensional. Numerically integrate this joint density over its 1st and 3rd components to obtain the marginal density $\widehat{f}_{\Pi_2(t_1,t_2)}$, an estimate of the pdf of linear combinations of the reduced form coefficients on Z_1.

2. (Convert to reduced form cf) Then note that

   $$\begin{aligned} \phi_{\pi_{12},\pi_{22}}(t_1, t_2) &= \mathbb{E}[\exp(i[t_1\pi_{12} + t_2\pi_{22}])] \\ &= \int_{\mathbb{R}} \exp(is) f_{\Pi_2(t_1,t_2)}(s)\, ds \end{aligned}$$

   and hence we can estimate the characteristic function of (π_{12}, π_{22}) by

   $$\widehat{\phi}_{\pi_{12},\pi_{22}}(t_1, t_2) = \int_{\mathbb{R}} \exp(is) \widehat{f}_{\Pi_2(t_1,t_2)}(s)\, ds,$$

   where numerical integration can be used to compute the integral.

3. (Convert to reduced form pdf) We now have the characteristic function of (π_{12}, π_{22}),

$$\begin{aligned}\phi_{\pi_{12},\pi_{22}}(t_1, t_2) &= \mathbb{E}[\exp(i[t_1\pi_{12} + t_2\pi_{22}])] \\ &= \int_{\mathbb{R}^2} \exp(i[t_1 p_1 + t_2 p_2]) f_{\pi_{12},\pi_{22}}(p_1, p_2)\ dp_1\ dp_2.\end{aligned}$$

Taking the inverse Fourier transform and substituting in our estimated characteristic function yields an estimator of the the joint pdf of (π_{12}, π_{22}):

$$\widehat{f}_{\pi_{12},\pi_{22}}(p_1, p_2) = \text{Re}\left[\frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} \exp(-i[t_1 p_1 + t_2 p_2])\widehat{\phi}_{\pi_{12},\pi_{22}}(t_1, t_2)\ dt_1\ dt_2\right].$$

Here $\text{Re}(z)$ stands for the real part of the complex number z. Again, we can use numerical integration to compute the integral, or the Fast Fourier Transform.

4. (Convert to structural pdf) Finally, note that

$$\gamma_2 = \frac{\pi_{22}}{\pi_{12}}.$$

Hence by theorem 3.1 of Curtiss (1941) we can write the density of γ_2 as

$$f_{\gamma_2}(z) = \int_{\mathbb{R}} |v| f_{\pi_{12},\pi_{22}}(v, zv)\ dv.$$

That is, the density of the ratio random variable depends on the integral of the joint density along a ray in $\mathbb{R}^2$ passing through the origin, whose slope is determined by z.

Taking sample analogs yields our final estimator:

$$\widehat{f}_{\gamma_2}(z) = \int_{\mathbb{R}} |v| \widehat{f}_{\pi_{12},\pi_{22}}(v, zv)\ dv,$$

where again we use numerical integration to compute the integral.

The first step involves estimating a single equation random coefficients model with exogenous regressors. There are several existing approaches for this. Beran and Hall's (1992) estimator requires all the coefficients to be independent, and hence cannot be used here. Beran and Millar (1994) consider a minimum distance estimator where the distribution of random coefficients is approximated by discretely supported distributions. Besides requiring numerical optimization, this approach produces an estimated $\widehat{f}_{\Pi_2(t_1,t_2)}$ which has discrete rather than continuous support, which may cause problems in steps 2–4 above. Instead, I recommend using one of the estimators proposed in Beran et al. (1996) and Hoderlein et al. (2010). Beran et al. (1996) propose to estimate the distribution of random coefficients by first estimating their characteristic function and then inverting it. Hoderlein et al. (2010) construct a regularized inverse Radon transform based kernel estimator; I use this estimator in my simulations and empirical application.

Both of the papers Beran et al. (1996) and Hoderlein et al. (2010) prove consistency and derive rates of convergence for their respective estimators, among other results. Consistency of $\widehat{f}_{\gamma_2}$ then follows since it is a sample analog estimator based on one of these first consistent first step estimators. I leave a full development of the asymptotic theory of $\widehat{f}_{\gamma_2}$ to future work. Finally, note that, beyond any necessary regularity conditions and those ensuring identification, the estimator described above does not restrict the joint distribution of unobservables.

While the above procedure can be extended immediately to allow for additional covariates X, this would involve estimating $1 + d_{Z_1} + d_{Z_2} + d_X$ dimensional joint density functions in the first step. One alternative is to assume that the coefficients δ_1 and δ_2 on the covariates are constant. For simplicity, consider the structural model (1) with $\delta_2 = 0$. Then the reduced form system is

$$\begin{aligned} Y_1 &= \frac{U_1 + \gamma_1 U_2}{1 - \gamma_1\gamma_2} + \frac{1}{1 - \gamma_1\gamma_2}(\delta_1' X) + \frac{\beta_1}{1 - \gamma_1\gamma_2} Z_1 + \frac{\gamma_1\beta_2}{1 - \gamma_1\gamma_2} Z_2 \\ Y_2 &= \frac{U_2 + \gamma_2 U_1}{1 - \gamma_1\gamma_2} + \frac{\gamma_2}{1 - \gamma_1\gamma_2}(\delta_1' X) + \frac{\gamma_2\beta_1}{1 - \gamma_1\gamma_2} Z_1 + \frac{\beta_2}{1 - \gamma_1\gamma_2} Z_2, \end{aligned}$$

which after defining some notation we write as

$$\begin{aligned} Y_1 &= \tilde{\pi}_{11} + \tilde{\pi}_{1x}(\delta_1' X) + \pi_{12} Z_1 + \pi_{13} Z_2 \\ Y_2 &= \tilde{\pi}_{21} + \tilde{\pi}_{2x}(\delta_1' X) + \pi_{22} Z_1 + \pi_{23} Z_2. \end{aligned}$$

If δ_1 was known, then this system would be the starting point for the estimator described above. In this case we could treat $\delta_1' X$ as a single scalar regressor, and hence we only have to estimate a 4 dimensional joint distribution instead of a $3 + d_X$ dimensional joint distribution. Since δ_1 is not known, this approach is not feasible. Instead, we can estimate

$$\tilde{\delta}_1 \equiv \mathbb{E}(\tilde{\pi}_{1x}\delta_1) = \mathbb{E}(\tilde{\pi}_{1x})\delta_1 = \mathbb{E}\left(\frac{1}{1 - \gamma_1\gamma_2}\right)(\delta_{11}, \ldots, \delta_{1K})'$$

by taking the coefficient on X in a linear mean regression of Y_1 on $(1, X, Z_1, Z_2)$. $\tilde{\delta}_1$ is not quite equal to δ_1 because of the $\mathbb{E}[1/(1 - \gamma_1\gamma_2)]$ scale factor. Nonetheless, we now have the system

$$\begin{aligned} Y_1 &= \tilde{\pi}_{11} + \frac{\tilde{\pi}_{1x}}{\mathbb{E}(\tilde{\pi}_{1x})}(\tilde{\delta}_1' X) + \pi_{12} Z_1 + \pi_{13} Z_2 \\ Y_2 &= \tilde{\pi}_{21} + \frac{\tilde{\pi}_{2x}}{\mathbb{E}(\tilde{\pi}_{1x})}(\tilde{\delta}_1' X) + \pi_{22} Z_1 + \pi_{23} Z_2. \end{aligned}$$

where the single index $\tilde{\delta}_1' X$ is estimated in the preliminary linear regression step. Thus, when estimating this system in step 1 by a single equation random coefficient estimator, we still obtain consistent estimates of the distribution of $t_1\pi_{12} + t_2\pi_{22}$ as needed.

For estimating single equation random coefficient models with many covariates, Hoderlein et al. (2010) proposed assuming δ_1 was constant, estimating it by a preliminary linear regression, and

then partialing it out as in partially linear models. This approach does not work here because the determinant term $1/(1-\gamma_1\gamma_2)$ ensures that all of the reduced form coefficients are random. Consequently, subtracting $\mathbb{E}(\tilde{\pi}_{1x}\delta_1')X$ from both sides of the reduced form equation does not remove the X term from the right hand side as it does in single equation models.

4.2 Monte Carlo simulations

To examine the nonparametric estimator's finite sample performance, I run several Monte Carlo simulations. The conditions of both theorems 2 and 3 hold in all simulations so that either result could be used to ensure identification. I consider four different data generating processes. They are identical along all dimensions except two. First, the common marginal distribution f_γ is one of the following:

1. f_γ is a truncated normal with pre-truncation mean 0.4 and standard deviation 0.05.
2. f_γ is a Beta distribution with shape parameter 6 and scale parameter 3.

See figure 2 for plots each of these marginal distributions. The support of the truncated normal and Beta is $[0,1]$, which is then scaled to $[0, 0.95]$, which helps ensure that f_γ is identified. Second, the instruments Z_1 and Z_2 are either standard Cauchy or $\mathcal{N}(0,3)$ distributed.

For each dgp I consider the sample sizes $N = 500$ and $N = 1000$. Both dgps have γ_1 independent of γ_2. Both dgps use the same distribution of additive unobservables (U_1, U_2), which are bivariate normal with $\mu_u = 0$, $\sigma_u = 1$, and $\rho_u = 0$. The instruments Z_1 and Z_2 have own coefficients $\beta_1 = 5$ and $\beta_2 = 0$, respectively, and friend coefficients 0 (e.g. the coefficient on Z_1 in the equation for Y_2 is zero), so that they satisfy the exclusion restriction. The constant term is -10. The true structural system with these parameter values is

$$\begin{aligned} Y_1 &= -10 + \gamma_1 Y_2 + 5Z_1 + 0Z_2 + U_1 \\ Y_2 &= -10 + \gamma_2 Y_1 + 0Z_1 + 5Z_2 + U_2. \end{aligned}$$

For each dgp, I compute several statistics. First, I compute the bias of several scalar parameter estimators. For any scalar parameter κ, the estimated bias is the mean of $\hat{\kappa}_s - \kappa$ over all $s = 1, \ldots, S$, where S is the total number of Monte Carlo simulations, and s indexes each simulation run. The estimated standard deviation is the standard deviation of $\hat{\kappa}_s - \kappa$ over all simulations s. The estimated MSE is the estimated bias squared plus the estimated standard deviation squared. I use $S = 250$ simulations, which yields simulation standard errors small enough to make statistically significant comparisons. I compute these statistics for the nonparametric estimator of the random coefficients' mean:

$$\widehat{\mathbb{E}}(\gamma) = \int_0^{0.95} x \cdot \widehat{f_\gamma}(x)\, dx,$$

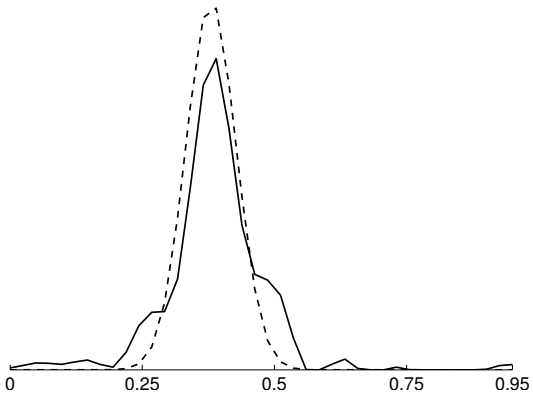
where $\widehat{f}_\gamma$ is the nonparametric estimator described earlier, as well as for the 2SLS estimator of the endogenous variable coefficient, viewed as an estimator of $\mathbb{E}(\gamma)$. I compute the estimated cdf of γ by

$$\widehat{F}_\gamma(t) = \int_0^t \widehat{f}_\gamma(x)\, dx$$

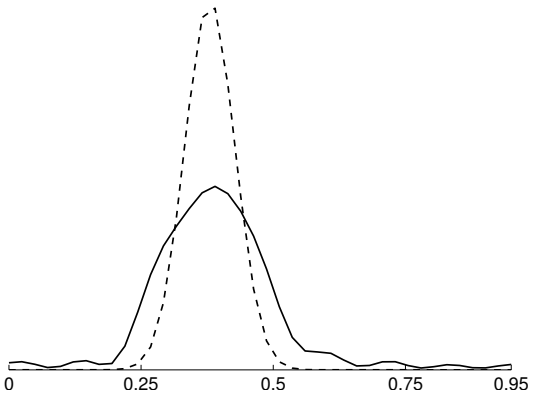
and use this to compute the estimated median $\widehat{\text{Med}}(\gamma)$ and interquartile range $\widehat{\text{IQR}}(\gamma)$. Finally, I compute the mean integrated squared error of the nonparametric estimator $\widehat{f}_\gamma$ of f_γ. For a fixed simulation s, the ISE is

$$\text{ISE}(\widehat{f}_{\gamma,s}) = \int_0^{0.95} [\widehat{f}_{\gamma,s}(x) - f_\gamma(x)]^2\, dx.$$

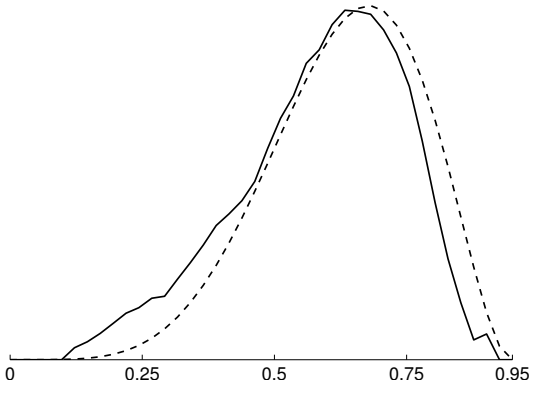
The mean ISE (MISE) is estimated by the mean of this value over all simulations.



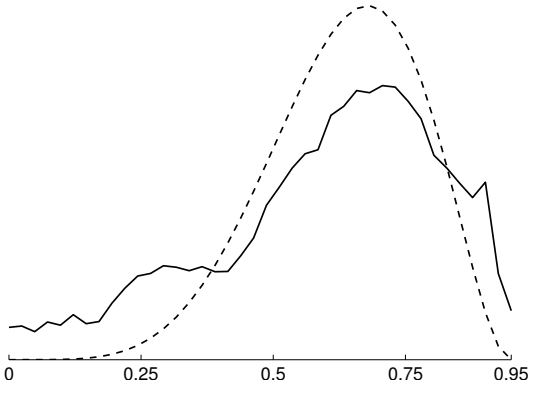
(a) Indep. trunc. normal(0.4,0.05), $N = 500$, Z_i standard Cauchy



(b) Indep. trunc. normal(0.4,0.05), $N = 500$, $Z_i \sim \mathcal{N}(0, 3)$



(c) Indep. Beta(6,3), $N = 500$, Z_i standard Cauchy



(d) Indep. Beta(6,3), $N = 500$, $Z_i \sim \mathcal{N}(0, 3)$

Figure 2: Nonparametric estimates of f_γ, the common marginal distribution of random coefficients. Dotted lines show the true density, solid lines show the estimated density. Estimates correspond to the simulation with integrated squared error at the median over all simulations.

Figure 2 shows example plots of $\widehat{f}_\gamma$ versus the true density, for $N = 500$. The estimator recovers the general shape of the true density in all four dgps, although it performs better with Cauchy distributed instruments compared to the normally distributed instruments. This is to be expected given the previous literature on nonparametric estimation in single equation random coefficients models. As discussed above, the only two options for the first step of my estimator are Beran et al. (1996) and Hoderlein et al. (2010). The assumptions in Beran et al. (1996) require thicker than normal tailed regressors. They also show that the rate of convergence depends on the rate at which the density of the regressors goes to zero in the tails: the thinner the regressor tails, the slower the rate. Likewise, the main theory of Hoderlein et al. (2010) also requires thicker than normal tailed regressors (see their theorem 3, however, where they show one way to relax this assumption). This property affects the first step of my estimator, and hence carries through to the final step estimator of $\widehat{f}_\gamma$, as we can see in the plots.

In addition to plotting the entire density of γ, we may also want to compute various summary statistics for this distribution. Tables 1 and 2 show the estimated bias, standard deviation, and MSE for the estimated mean $\widehat{\mathbb{E}}(\gamma)$, median $\widehat{\text{Med}}(\gamma)$, and interquartile range $\widehat{\text{IQR}}(\gamma)$, all obtained from $\widehat{f}_\gamma$. I call these the RC estimators. For each dgp, the true values of these parameters are also shown. For comparison, I also show the estimated bias, standard deviation, and MSE of the 2SLS estimator, viewed as an estimator of $\mathbb{E}(\gamma)$, although recall that the 2SLS estimand is generally not equal to the mean random coefficient (see section 2.3). Finally, I also show the mean ISE and the standard deviation of the ISE. Table 1 shows results for Cauchy distributed instruments, while table 2 shows results for normally distributed instruments.

Table 1: Monte Carlo results: Cauchy Z

	$\widehat{\mathbb{E}(\gamma)}$		$\widehat{\text{Med}}(\gamma)$	$\widehat{\text{IQR}}(\gamma)$	MISE
	2SLS	RC	RC	RC	RC
Indep. trunc. normal(0.4,0.05)	$\mathbb{E}(\gamma) = 0.38$		$\text{Med}(\gamma) = 0.38$	$\text{IQR} = 0.0641$	
$N = 500$	-0.0007	0.0031	0.0012	0.0253	0.4763
	[0.0317]	[0.0061]	[0.0068]	[0.0108]	[0.2634]
	(0.0010)	(0.0000)	(0.0000)	(0.0008)	
$N = 1000$	0.0003	0.0031	0.0012	0.0231	0.4037
	[0.0368]	[0.0046]	[0.0051]	[0.0082]	[0.1937]
	(0.0014)	(0.0000)	(0.0000)	(0.0006)	
Indep. Beta(6,3)	$\mathbb{E}(\gamma) = 0.63$		$\text{Med}(\gamma) = 0.6455$	$\text{IQR} = 0.202$	
$N = 500$	0.0116	-0.0438	-0.0344	0.0171	0.0841
	[0.0966]	[0.0207]	[0.0196]	[0.0202]	[0.0593]
	(0.0095)	(0.0023)	(0.0016)	(0.0007)	
$N = 1000$	0.0082	-0.0455	-0.0362	0.0142	0.0779
	[0.0961]	[0.0165]	[0.0150]	[0.0179]	[0.0431]
	(0.0093)	(0.0023)	(0.0015)	(0.0005)	

For each dgp: Bias is first. Standard deviations in brackets. MSE in parentheses.

First consider table 1, with Cauchy distributed instruments. The first dgp is similar to a model with a constant coefficient of 0.38. It is symmetric around 0.38 with all the mass within $[0.25, 0.5]$. Both the RC and the 2SLS estimator estimate $\mathbb{E}(\gamma)$ well, although the standard deviation of 2SLS is substantially larger than the RC estimator. The RC estimator of the median similarly performs well. The RC IQR estimator is biased upwards by about 33%, which can be seen in figure 2, since the estimated pdf is more spread out than the truth.

The second dgp is slightly asymmetric and more spread out than the first dgp. In this case, both estimators do worse than in the first dgp in estimating the mean. While 2SLS has a smaller bias then the RC estimator, 2SLS again has a substantially larger standard deviation, which implies that the RC estimator's MSE is four times smaller than that of 2SLS. The RC median estimator has a smaller bias than the RC mean estimator. The RC IQR estimator performs well in this dgp, with a bias and standard deviation one order of magnitude smaller than the truth.

Table 2: Monte Carlo results: Normal Z

	$\widehat{\mathbb{E}(\gamma)}$		$\widehat{\text{Med}}(\gamma)$	$\widehat{\text{IQR}}(\gamma)$	MISE
	2SLS	RC	RC	RC	RC
Indep. trunc. normal(0.4,0.05)	$\mathbb{E}(\gamma) = 0.38$		$\text{Med}(\gamma) = 0.38$	$\text{IQR} = 0.0641$	
$N = 500$	0.0017	0.0055	0.0041	0.0638	1.3836
	[0.0051]	[0.0069]	[0.0057]	[0.0103]	[0.2810]
	(0.0000)	(0.0001)	(0.0000)	(0.0042)	
$N = 1000$	0.0010	0.0049	0.0034	0.0644	1.3934
	[0.0034]	[0.0053]	[0.0042]	[0.0080]	[0.2151]
	(0.0000)	(0.0001)	(0.0000)	(0.0042)	
Indep. Beta(6,3)	$\mathbb{E}(\gamma) = 0.63$		$\text{Med}(\gamma) = 0.6455$	$\text{IQR} = 0.202$	
$N = 500$	0.0223	-0.0435	-0.0188	0.0920	0.1898
	[0.0137]	[0.0109]	[0.0122]	[0.0160]	[0.0446]
	(0.0007)	(0.0020)	(0.0005)	(0.0087)	
$N = 1000$	0.0237	-0.0428	-0.0178	0.0908	0.1850
	[0.0099]	[0.0072]	[0.0083]	[0.0123]	[0.0350]
	(0.0007)	(0.0019)	(0.0004)	(0.0084)	

For each dgp: Bias is first. Standard deviations in brackets. MSE in parentheses.

Next consider table 2, with normally distributed instruments. Consider the first dgp. Despite the problems mentioned earlier with relatively thin tailed regressors, the RC estimators of the mean and median do very well. The RC estimators of the location of the distribution are comparable to 2SLS, which now also performs well with both a small bias and a small standard deviation. The RC IQR estimator performs worse. It is two times larger than the true IQR on average. This can also be seen in figure 2. In the second dgp, again both estimators do worse than the first dgp in estimating the location of the distribution. The RC estimator of the mean and 2SLS are comparable, while the RC median estimator performs better than both. The RC IQR estimator is now overshooting the truth by about 45% on average.

Overall, the simulation results suggest that the RC estimator performs well with practical sample sizes. In addition to providing good estimators of the center of the distribution, it provides reasonable estimators of the spread, and of the entire shape of the distribution. In contrast, traditional analysis based on the 2SLS estimand necessarily provides a limited summary of the distribution of γ.

4.3 Bandwidth selection

The first step inverse Radon transform estimator requires choosing a bandwidth. In the Monte Carlo simulations, I follow Hoderlein et al. (2010) and minimize the mean density weighted ISE,

$$\mathbb{E}\left[\int_0^{0.95} [\widehat{f_\gamma}(x) - f_\gamma(x)]^2 f_\gamma(x)\, dx\right].$$

Since computing this number requires knowledge of the true density f_γ, this approach is not feasible in practice. As of now, there do not exist any data-based methods for choosing the bandwidth when estimating single equation random coefficient models, for either the inverse Radon transform estimator of Hoderlein et al. (2010) or the characteristic function inversion estimator of Beran et al. (1996). It is likely that reasonable methods, such as plug-in, resampling, or cross-validation based approaches, can be developed by following the related problem of bandwidth selection in measurement error deconvolution estimators, for example. Developing such methods is beyond the scope of the present paper. Instead, for choosing the bandwidth in my empirical application, I propose the following first pass approach.

First, notice that in step 3 of the RC estimator we need to take an integral over (t_1, t_2). For this step, in both the simulations and empirical illustration, I use a 1000 point Halton grid. For each of these grid points, we have to compute the first step single equation estimator. Hence there are potentially 1000 different bandwidths we must choose, corresponding to the different values of (t_1, t_2) in our grid. For any given point in the (t_1, t_2) grid, we can choose the bandwidth by visually inspecting the plot of $f_{\Pi_2(t_1,t_2)}$. Even in the related problem of measurement error deconvolution, where several data-driven bandwidth estimators actually do exist, some authors prefer this visual method; see Carroll, Ruppert, Stefanski, and Crainiceanu (2006) page 283. The problem is that we cannot practically do this manually 1000 different times. Instead, I pick a single bandwidth visually, and then scale it up or down automatically according to the range of the support of $t_1\pi_{12} + t_2\pi_{22}$, which depends on the values of t_1 and t_2.

To see why this is a reasonable first pass method for choosing all of the bandwidths simultaneously, consider the standard problem of estimating the density of a random variable X. Let h be an optimally chosen bandwidth for estimating f_X. Then ah will be the optimal bandwidth for estimating the density f_{aX} of the scaled random variable aX, for $a \neq 0$. This is the same idea I use here. The analogy to estimating f_{aX} is not quite right, because we're taking a linear combination of two dependent random variables, rather than just estimating a single random variable. Nonetheless, by visually inspecting the plots $f_{\Pi_2(t_1,t_2)}$ for various (t_1, t_2), this method seems to work reasonably well.

5 Empirical illustration: Peer effects in education

In this section, I illustrate how to use the methods developed in this paper by exploring heterogeneous peer effects in education. Sacerdote (2011) and Epple and Romano (2011) give extensive surveys of this literature. I construct pairs of best friends using the Add Health dataset (Harris, Halpern, Whitsel, Hussey, Tabor, Entzel, and Udry 2009). I then apply the kernel estimator described in section 4.1 to nonparametrically estimate the distribution of random coefficients γ_1 and γ_2 in the simultaneous equations model (1), where outcomes are high school GPAs. Following one specification in Sacerdote (2000, 2001), I use lagged outcomes as instruments. My approach yields estimates of the average endogenous social effect, as well as other distributional features like quantile endogenous social effects, while allowing that not all people affect their best friend equally.

5.1 The Add Health dataset

Add Health is a panel dataset of students who were in grades 7-12 in the United States during the 1994 to 1995 school year. There have been four completed waves of data collection. I use data from the wave 1 in-home survey, administered between April and December 1995. In this survey, students were asked to name up to 5 male friends and up to 5 female friends. These friendship data have been widely used to study the impact of social interactions on many different outcomes of interest (e.g., Bramoullé et al. 2009 and Lin 2010). Card and Giuliano (2013) use this friendship data to construct pairs of best friends. They then study social interaction effects on risky behavior, such as smoking and sexual activity, by estimating discrete game models. These are simultaneous equations models with discrete outcomes and two equations, where each equation represents one friend's best-response function of the other friend's action. I follow a similar approach, but with continuous outcomes and allowing for nonparametric heterogeneous social effects.

I also use data from the Adolescent Health and Academic Achievement (AHAA) study, which obtained transcript release forms from students during the Add Health wave 3 survey administered between 2001 and 2002. AHAA linked detailed high school transcript data with the earlier surveys. 12,237 students are in the AHAA study. Among these students, I keep only students in grades 10–12 (or higher, due to repeated grades) during the wave 1 survey school year, 1994–1995. Middle schoolers and 9th graders get dropped because AHAA only collected high school transcript data and hence I do not have lagged GPAs for them. This leaves 6,585 students. Another 60 students get dropped due to missing contemporaneous or lagged GPA data, leaving 6,525 students. From these students, I construct 330 same-sex pairs of students—660 students total. Students were asked to list their top 5 friends starting with their first best friend, and then their second best friend, and so on. I first pair all students who named each other as their first best friend. I then pair students where one student was named as a best friend, but the other student was only named as a second best friend. I next pair students where both students named each other as second best friends, and so on. Note that no student is included more than once. The overall sample size is relatively small

because in order to enter the final sample both students in the pair had to be among the 6,525 students from the AHAA sample of 10–12th graders. If a student named friends who were in 9th grade or middle school, or who were not even in the original Add Health sample (90,118 students from in-school wave 1), then that student does not appear in my final sample.

5.2 Empirical results

I estimate a random coefficients analog of equations (8) and (9) in Sacerdote (2000),

$$\begin{aligned} \text{GPA}_{1,t} &= \gamma_1 \text{GPA}_{2,t} + \beta_1 \text{GPA}_{1,t-1} + U_{1,t} \\ \text{GPA}_{2,t} &= \gamma_2 \text{GPA}_{1,t} + \beta_2 \text{GPA}_{2,t-1} + U_{2,t}. \end{aligned} \tag{9}$$

Here the outcome of interest is a student's GPA during the 1994–1995 school year. The explanatory variables are their best friend's contemporaneous GPA, and their own GPA in the previous school year. Table **??** shows summary statistics; there is substantial variation in both current and lagged GPA. System (9) is a special case of equations (1) and (2) in Sacerdote (2001), where we assume no measurement error in lagged outcomes and no contextual effect of your best friend's lagged outcomes. As in Sacerdote (2001), controlling for lagged outcomes is viewed as a way to condition on ability. Consequently, the exclusion restriction here says that your best friend's ability does not directly affect your performance this year. Instead, specification (9) only allows your best friend's contemporaneous studies and effort to affect your GPA.

Table 3: Summary statistics

	count	p50	mean	sd	min	max
Current GPA	660	2.9	2.74	0.90	.08	4
Lagged GPA	660	2.9	2.83	0.82	.11	4

Besides exclusion, the next assumption needed to apply an instrumental variable identification strategy is exogeneity. Here that requires your best friend's past performance to be unrelated to all unobserved factors that affect your current performance, including your random coefficients. Given that friendships likely form nonrandomly, this is perhaps implausible in the current setting. Nonetheless, similar assumptions have been used in previous research with the Add Health data, like Card and Giuliano (2013). Moreover, this assumption is often plausible in other datasets, to which my methods would apply. For example, in Sacerdote's original data roommates were matched randomly, which he argues justifies the exogeneity assumption.

The final assumptions needed to apply the identification result theorem 2 of section 3 are continuity of the instrument, which holds here because GPA is a continuous variable, and relevance—your past GPA must affect your current GPA. Table 4 shows estimates of the reduced form equations of current GPA on own and friend's lagged GPA. They are obtained via SUR under the restriction that the coefficients on own and friend GPA are equal across equations. This constraint holds

because labels of friend 1 versus friend 2 are arbitrary. This constraint holds regardless of whether the coefficients are constant or random. Moreover, since the reduced form equations only contain exogenous regressors, the SUR estimates are consistent for the mean reduced form random coefficients. Own lagged GPA has a large positive effect on own current GPA, suggesting that the relevance assumption holds.

Table 4: Reduced form regression

	Own current GPA
Own lagged GPA	0.8167
	[0.7651, 0.8683]
Friend's lagged GPA	0.1512
	[0.0997, 0.2027]
R^2	0.65
Observations	330

Observations are pairs of best friends. 95% confidence intervals shown in brackets. Estimates obtained from SUR with cross-equation constraints; see body text for details.

Table 5: Estimates of endogenous social interaction effect

	SUR	3SLS	RC
$\widehat{\mathbb{E}}(\gamma)$	0.2965	0.1859	0.5383
	[0.2477,0.3453]	[0.1196,0.2522]	[0.5249,0.6384]
$Q_\gamma(0.25)$			0.3300
			[0.3267,0.4496]
$\text{Med}(\gamma)$			0.6457
			[0.6281,0.7101]
$Q_\gamma(0.75)$			0.7199
			[0.7177,0.8330]
Observations	330	330	330

Observations are pairs of best friends. 95% confidence intervals shown in brackets. See body text for details of estimation.

Table 5 shows the main estimation results. First, SUR provides estimates of system (9), ignoring the simultaneity problem, and imposing the constraint that the coefficients on each equation

are equal ($\gamma_1 = \gamma_2$, $\beta_1 = \beta_2$), as discussed earlier. This gives a single point estimate of the coefficient on friend's GPA, shown in the first row of the table. These estimates describe the correlation between peer outcomes. Next, 3SLS provides estimates of system (9), also with the cross-equation constraints, but using friend's lagged GPA as instruments. The 3SLS point estimate of the endogenous social interaction effect implies that a one point increase in your friend's GPA increases your own GPA by about 0.19 points, with a 95% confidence interval of $[0.12, 0.25]$. As discussed earlier, when the endogenous variables have random coefficients, estimators like 2SLS and 3SLS estimate weighted average effects, not the mean of the random coefficients. Moreover, these estimates can be quite different from the actual average coefficient. The RC estimator described in section 4.1, on the other hand, provides a consistent estimator of the average random coefficient, as well as their distribution.

Because the labels of friend 1 versus friend 2 are arbitrary, the marginal distributions of γ_1 and γ_2 are equal, $f_{\gamma_1} = f_{\gamma_2}$. I estimate this common marginal by applying the RC estimator to both γ_2 and γ_1 and then averaging the two estimators: $\widehat{f}_\gamma = (\widehat{f}_{\gamma_1} + \widehat{f}_{\gamma_2})/2$. (These two estimators separately look quite similar.) Using this estimated marginal distribution, I compute the mean, 25th percentile, median, and 75th percentile of the distribution of endogenous social interaction effects. These estimates are shown in the third column of table 5. 95% confidence intervals are shown in brackets, using the bootstrap percentile method with 250 bootstrap samples. The mean estimate is comparable to the 3SLS estimates, in the sense that they are asymptotically equal under constant coefficients.

These estimates suggest two things: First, there is substantial heterogeneity in the distribution of endogenous social effects. Second, the unweighted average effect is higher than the 3SLS estimand, whose point estimate is about 0.19. Recall from section 2.3 that the 2SLS estimand for equation 1 is a weighted average of γ_1, where the weights depend on the strength of the instrument (your friend's lagged GPA) and how close your system is to being parallel (the size of the determinant term $1 - \gamma_1\gamma_2$). Hence the 2SLS estimand can be smaller than the true average coefficient for several reasons. For example, suppose people who are not too socially susceptible (small γ_1) are more likely to be friends with people whose current academic performance depends strongly on their past academic performance (large β_2). This will tend to make the 2SLS estimand smaller than the unweighted average random coefficient. While I am unaware of similar derivations in the literature for the constrained 3SLS estimand, it is likely to have a similar interpretation as 2SLS.

While functionals like the mean and quantiles are usually estimated much more precisely than entire functions, it can still be informative to examine the overall shape of the estimated density of γ. Figure 3 plots this estimated density. There are two distinct groups. About 40% of people have endogenous social interaction effects between 0 and 0.4 while about 55% of people are between 0.55 and 0.8. In this case, the density itself is informative above and beyond the mean and the quartiles.

Overall, these results suggest that for many students, social influence matters for high school GPA, which is consistent with the existing empirical literature. The RC estimated distribution

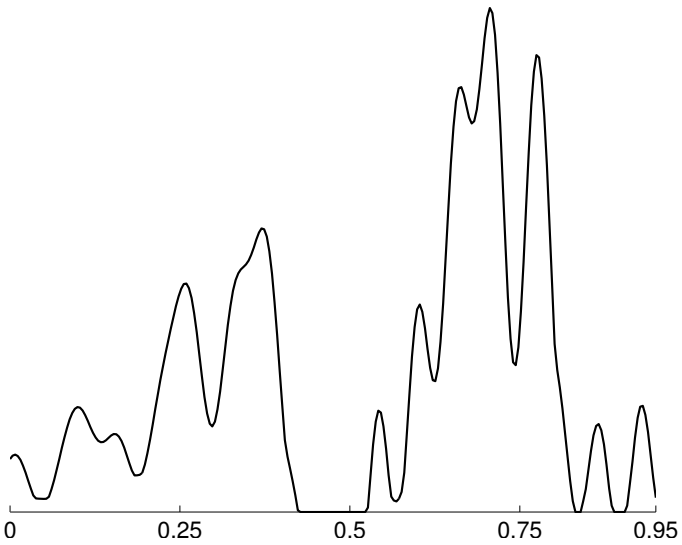


Figure 3: Nonparametric estimate of the density of endogenous social interaction effects.

suggests that there is substantial heterogeneity in social influence, with roughly half of students being strongly influenced by their best friend and another half still being influenced, but to a much smaller extent. Moreover, within both of these groups the average effect exceeds the 3SLS point estimate. This suggests that, when examining peer effects on GPA in high school, findings of social interaction effects based on 2SLS or 3SLS may understate potential multiplier effects of policy interventions.

In this section I have illustrated how to use the methods developed in this paper in practice. For this reason I have focused on a clearly simplified setup and specification. Further analysis would include estimating distributions of social interaction effects conditional on covariates, which may help explain the observed bimodality of effects. Such analysis may reveal which covariate combinations lead to large average effects. This, in turn, may help policy makers choose which students to target for interventions. More generally, I hope that the methods in this paper will help researchers understand, identify, and estimate unobserved heterogeneity in various applied settings with simultaneity.

6 Conclusion

In this paper I have studied identification of linear simultaneous equations models with random coefficients. In simultaneous systems, random coefficients on endogenous variables pose qualitatively different problems from random coefficients on exogenous variables. The possibility of nearly parallel lines can cause classical mean-based identification approaches to fail. For systems of two equations, I showed that, even allowing for nearly parallel lines, we can still identify the marginal distributions of random coefficients by using a full support instrument. When nearly parallel lines are ruled out, we can relax the full support assumption. I proposed a consistent nonparametric

estimator for the distribution of coefficients, and show that it performs well in finite samples. I applied my results to analyze peer effects in educational achievement and found evidence of significant heterogeneity as well as mean coefficient estimates larger than the usual point estimates.

Several issues remain for future research. First, several estimation issues remain, such as a full analysis of inference for the nonparametric estimator. Second, I have shown that although the full joint distribution of structural unobservables is not point identified, some marginal distributions are point identified. It would be helpful to have a complete characterization of the identified set for the joint distribution of structural unobservables.

References

Angrist, J. D. (2004): "Treatment effect heterogeneity in theory and practice," *The Economic Journal*, 114, C52–C83.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economic Studies*, 67, 499–527.

Angrist, J. D. and G. W. Imbens (1995): "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American Statistical Association*, 90, 431–442.

Arellano, M. and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models," *Review of Economic Studies*, 79, 987–1020.

Bayer, C. and J. Teichmann (2006): "The proof of Tchakaloff's theorem," *Proceedings of the American Mathematical Society*, 134, 3035–3040.

Bélisle, C., J.-C. Massé, and T. Ransford (1997): "When is a probability measure determined by infinitely many projections?" *The Annals of Probability*, 767–786.

Benkard, C. and S. Berry (2006): "On the nonparametric identification of nonlinear simultaneous equations models: Comment on Brown (1983) and Roehrig (1988)," *Econometrica*, 74, 1429–1440.

Beran, R. (1995): "Prediction in random coefficient regression," *Journal of Statistical Planning and Inference*, 43, 205–213.

Beran, R., A. Feuerverger, and P. Hall (1996): "On nonparametric estimation of intercept and slope distributions in random coefficient regression," *Annals of Statistics*, 24, 2569–2592.

Beran, R. and P. Hall (1992): "Estimating coefficient distributions in random coefficient regressions," *Annals of Statistics*, 20, 1970–1984.

Beran, R. and P. Millar (1994): "Minimum distance estimation in random coefficient regression models," *Annals of Statistics*, 22, 1976–1992.

Berry, S. and P. Haile (2011): "Identification in a class of nonparametric simultaneous equations models," *Working paper*.

Berry, S. T. and P. A. Haile (2014): "Identification in differentiated products markets using market level data," *Econometrica*, 82, 1749–1797.

Bjorn, P. and Q. Vuong (1984): "Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation," *Working paper*.

Blume, L. E., W. A. Brock, S. N. Durlauf, and Y. M. Ioannides (2011): "Identification of social interactions," *Handbook of Social Economics*, 1, 853–964.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015): "Linear social network models," *Journal of Political Economy*, 123, 444–496.

Blundell, R. and R. L. Matzkin (2014): "Control functions in nonseparable simultaneous equations models," *Quantitative Economics*, 5, 271–295.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009): "Identification of peer effects through social networks," *Journal of Econometrics*, 150, 41–55.

Bramoullé, Y. and R. Kranton (2015): "Games played on networks," *Working paper*.

Bramoullé, Y., R. Kranton, and M. D'Amours (2014): "Strategic interaction and networks," *The American Economic Review*, 104, 898–930.

Bresnahan, T. and P. Reiss (1991): "Empirical models of discrete games," *Journal of Econometrics*, 48, 57–81.

Brock, W. A. and S. N. Durlauf (2001): "Interactions-based models," *Handbook of Econometrics*, 5, 3297–3380.

Brown, B. (1983): "The identification problem in systems nonlinear in the variables," *Econometrica*, 51, 175–196.

Browning, M., P.-A. Chiappori, and Y. Weiss (2014): *Economics of the family*, Cambridge University Press.

Card, D. and L. Giuliano (2013): "Peer effects and multiple equilibria in the risky behavior of friends," *Review of Economics and Statistics*, 95, 1130–1149.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006): *Measurement error in nonlinear models: A modern perspective*, CRC press.

Case, A. (1991): "Spatial patterns in household demand," *Econometrica*, 59, 953–965.

Chao, J. C. and P. C. Phillips (1998): "Posterior distributions in limited information analysis of the simultaneous equations model using the Jeffreys prior," *Journal of Econometrics*, 87, 49–86.

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261.

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405–1441.

——— (2009): "Excess heterogeneity, endogeneity and index restrictions," *Journal of Econometrics*, 152, 37–45.

Christakis, N. and J. Fowler (2007): "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, 357, 370–379.

Cohen-Cole, E. and J. Fletcher (2008): "Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic," *Journal of Health Economics*, 27, 1382–1387.

Cramér, H. and H. Wold (1936): "Some theorems on distribution functions," *Journal of the London Mathematical Society*, 1, 290–294.

Cuesta-Albertos, J., R. Fraiman, and T. Ransford (2007): "A sharp form of the Cramér–Wold theorem," *Journal of Theoretical Probability*, 20, 201–209.

Curtiss, J. (1941): "On the distribution of the quotient of two chance variables," *Annals of Mathematical Statistics*, 12, 409–421.

Duflo, E. and E. Saez (2003): "The role of information and social interactions in retirement plan decisions: evidence from a randomized experiment," *The Quarterly Journal of Economics*, 118, 815–842.

Dunker, F., S. Hoderlein, and H. Kaido (2013): "Random coefficients in static games of complete information," *Working paper*.

Elaydi, S. (2005): *An introduction to difference equations*, Springer, third ed.

Epple, D. and R. Romano (2011): "Peer effects in education: A survey of the theory and evidence," *Handbook of Social Economics*, 1, 1053–1163.

Evans, W., W. Oates, and R. Schwab (1992): "Measuring peer group effects: A study of teenage behavior," *Journal of Political Economy*, 966–991.

Falk, A. and A. Ichino (2006): "Clean evidence on peer effects," *Journal of Labor Economics*, 24, 39–57.

Fisher, F. M. (1966): *The identification problem in econometrics*, McGraw-Hill.

Fox, J. T. and A. Gandhi (2011): "Identifying demand with multidimensional unobservables: a random functions approach," *Working paper*.

Fox, J. T., K. Kim, S. P. Ryan, and P. Bajari (2012): "The random coefficients logit model is identified," *Journal of Econometrics*, 166, 204–212.

Fox, J. T. and N. Lazzati (2013): "Identification of discrete choice models for bundles and binary games," *Working paper*.

Gautier, E. and S. Hoderlein (2012): "A triangular treatment effect model with random coefficients in the selection equation," *Working paper*.

Gautier, E. and Y. Kitamura (2013): "Nonparametric estimation in random coefficients binary choice models," *Econometrica*, 81, 581–607.

Graham, B. S. and J. L. Powell (2012): "Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models," *Econometrica*, 80, 2105–2152.

Hahn, J. (2001): "Consistent estimation of the random structural coefficient distribution from the linear simultaneous equations system," *Economics Letters*, 73, 227–231.

Harris, K., C. Halpern, E. Whitsel, J. Hussey, J. Tabor, P. Entzel, and J. Udry (2009): "The national longitudinal study of adolescent health: research design," *WWW document*.

Hausman, J. A. (1983): "Specification and estimation of simultaneous equation models," *Handbook of Econometrics*, 391–448.

Heckman, J. J., D. Schmierer, and S. Urzua (2010): "Testing the correlated random coefficient model," *Journal of Econometrics*, 158, 177–203.

Heckman, J. J. and E. J. Vytlacil (1998): "Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling," *Journal of Human Resources*, 33, 974–987.

——— (2007): "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments," *Handbook of Econometrics*, 6.

Hildreth, C. and J. Houck (1968): "Some estimators for a linear model with random coefficients," *Journal of the American Statistical Association*, 63, 584–595.

Hirano, K. and J. Hahn (2010): "Design of randomized experiments to measure social interaction effects," *Economics Letters*, 106, 51–53.

Hoderlein, S., J. Klemelä, and E. Mammen (2010): "Analyzing the random coefficient model nonparametrically," *Econometric Theory*, 26, 804–837.

Hoderlein, S. and E. Mammen (2007): "Identification of marginal effects in nonseparable models without monotonicity," *Econometrica*, 75, 1513–1518.

Hoderlein, S., L. Nesheim, and A. Simoni (2012): "Semiparametric estimation of random coefficients in structural economic models," *Working paper*.

Hoderlein, S. and R. Sherman (2013): "Identification and estimation in a correlated random coefficients binary response model," *Working paper*.

Horn, R. A. and C. R. Johnson (2013): *Matrix Analysis*, Cambridge University Press, second ed.

Horowitz, J. L. and C. F. Manski (1995): "Identification and robustness with contaminated and corrupted data," *Econometrica*, 63, 281–302.

Hsiao, C. (1983): "Identification," *Handbook of Econometrics*, 1, 223–283.

Hsiao, C. and M. Pesaran (2008): "Random coefficient models," in *The Econometrics of Panel Data*, ed. by L. Mátyás and P. Sevestre, Springer-Verlag, vol. 46 of *Advanced Studies in Theoretical and Applied Econometrics*, chap. 6, 185–213, third ed.

Ichimura, H. and T. S. Thompson (1998): "Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution," *Journal of Econometrics*, 86, 269–295.

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

Intriligator, M. (1983): "Economic and econometric models," *Handbook of Econometrics*, 1, 181–221.

Kasy, M. (2014): "Instrumental variables with unrestricted heterogeneity and continuous treatment," *Review of Economic Studies*, 81, 1614–1636.

Kelejian, H. (1974): "Random parameters in a simultaneous equation framework: identification and estimation," *Econometrica*, 42, 517–527.

Kleibergen, F. and H. K. van Dijk (1994): "Bayesian analysis of simultaneous equation models using noninformative priors," *Tinbergen Institution Discussion Paper TI94-134*.

Landsberg, J. M. (2012): *Tensors: geometry and applications*, American Mathematical Society.

Lee, L.-F., X. Liu, and X. Lin (2010): "Specification and estimation of social interaction models with network structures," *Econometrics Journal*, 13, 145–176.

Lewbel, A. (2007): "Coherency and completeness of structural models containing a dummy endogenous variable," *International Economic Review*, 48, 1379–1392.

Lin, X. (2010): "Identifying peer effects in student academic achievement by spatial autoregressive models with group unobservables," *Journal of Labor Economics*, 28, 825–860.

Manski, C. F. (1993): "Identification of endogenous social effects: the reflection problem," *Review of Economic Studies*, 60, 531–542.

——— (1995): *Identification problems in the social sciences*, Cambridge: Harvard University Press.

——— (1997): "Monotone Treatment Response," *Econometrica*, 65, 1311–1334.

Matzkin, R. L. (2003): "Nonparametric estimation of nonadditive random functions," *Econometrica*, 71, 1339–1375.

——— (2007): "Nonparametric identification," *Handbook of Econometrics*, 6, 5307–5368.

——— (2008): "Identification in nonparametric simultaneous equations models," *Econometrica*, 76, 945–978.

——— (2012): "Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity," *Journal of Econometrics*, 166, 106–115.

Moffitt, R. A. (2001): "Policy interventions, low-level equilibria, and social interactions," *Social dynamics*, 4, 45–82.

Munkres, J. R. (1991): *Analysis on manifolds*, Westview Press.

Okamoto, M. (1973): "Distinctness of the eigenvalues of a quadratic form in a multivariate sample," *The Annals of Statistics*, 763–765.

Okumura, T. (2011): "Nonparametric estimation of labor supply and demand factors," *Journal of Business & Economic Statistics*, 29, 174–185.

Petersen, L. C. (1982): "On the relation between the multidimensional moment problem and the one-dimensional moment problem," *Mathematica Scandinavica*, 51, 361–366.

Ponomareva, M. (2010): "Quantile regression for panel data models with fixed effects and small T: Identification and estimation," *Working paper*.

Raj, B. and A. Ullah (1981): *Econometrics: A varying coefficients approach*, Croom Helm.

Robert, C. (1991): "Generalized inverse normal distributions," *Statistics & Probability Letters*, 11, 37–41.

Roehrig, C. (1988): "Conditions for identification in nonparametric and parametric models," *Econometrica*, 56, 433–447.

Rossi, H. and R. C. Gunning (1965): *Analytic functions of several complex variables*, Prentice-Hall, Inc.

Rubin, H. (1950): "Note on random coefficients," in *Statistical lnference in Dynamic Economic Models*, ed. by T. C. Koopmans, John Wiley & Sons, Inc. New York, vol. 10 of *Cowles Commission Monographs*, 419–421.

Sacerdote, B. (2000): "Peer effects with random assignment: Results for Dartmouth roommates," *NBER Working Paper*.

——— (2001): "Peer effects with random assignment: results for Dartmouth roommates," *The Quarterly Journal of Economics*, 116, 681–704.

——— (2011): "Peer effects in education: How might they work, how big are they and how much do we know thus far?" *Handbook of the Economics of Education*, 3, 249–277.

Swamy, P. (1968): "Statistical inference in random coefficient regression models," Ph.D. thesis, University of Wisconsin–Madison.

——— (1970): "Efficient inference in a random coefficient regression model," *Econometrica*, 38, 311–323.

Tamer, E. (2003): "Incomplete simultaneous discrete response model with multiple equilibria," *Review of Economic Studies*, 70, 147–165.

Torgovitsky, A. (2014): "Identification of nonseparable models using instruments with small support," *Econometrica, Forthcoming*.

Wooldridge, J. M. (1997): "On two stage least squares estimation of the average treatment effect in a random coefficient model," *Economics Letters*, 56, 129–133.

——— (2003): "Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model," *Economics Letters*, 79, 185–191.

Data References

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University

of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

This research uses data from the AHAA study, which was funded by a grant (R01 HD040428-02, Chandra Muller, PI) from the National Institute of Child Health and Human Development, and a grant (REC-0126167, Chandra Muller, PI, and Pedro Reyes, Co-PI) from the National Science Foundation. This research was also supported by grant, 5 R24 HD042849, Population Research Center, awarded to the Population Research Center at The University of Texas at Austin by the Eunice Kennedy Shriver National Institute of Health and Child Development. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

A Proofs

Remark 3. Kelejian's (1974) condition for identification is that $\det(I-\Gamma)$ does not depend on the random components of Γ. In the two equation system $\det(I-\Gamma) = 1-\gamma_1\gamma_2$. So his results apply if either γ_1 or γ_2 is zero with probability one; that is, if system (1) is actually triangular, and there is no feedback between Y_1 and Y_2. □

Remark 4. Hahn's (2001) identification result, his lemma 1, applies Beran and Millar (1994) proposition 2.2. As discussed in section 3.2 on SUR models, the assumptions in that proposition rule out common regressors, which in turn rules out fully simultaneous equations models, as well as triangular models, as discussed in section 3.3. More specifically, consider system (1) with no covariates X for simplicity. In this model, Hahn's support condition (assumption v) assumes the support of $t_1 + t_2Z_1 + t_3Z_2$ contains an open ball in $\mathbb{R}$ for all nonzero $(t_1, t_2, t_3) \in \mathbb{R}^3$. Beran and Millar's support condition is that the support of $(t_1Z_1,\ t_1Z_2,\ t_2Z_1,\ t_2Z_2)$ contains an open ball in $\mathbb{R}^4$ for all $(t_1, t_2) \in \mathbb{R}^2$, t_1, t_2 nonzero. Hahn's condition is not sufficient for Beran and Millar's, but for the reasons disucssed in sections 3.2 and 3.3, Beran and Millar's condition cannot hold in system (1) regardless. Thus neither the results of Beran and Millar (1994) nor those of Hahn (2001) apply to the fully simultaneous equations model considered here, or even to triangular models. □

Derivations to show 2SLS estimates a weighted average effect parameter. We have

$$\begin{aligned}
\operatorname{cov}(Y_1, Z_2) &= \mathbb{E}[(\gamma_1 Y_2 + U_1)(Z_2 - \mathbb{E}(Z_2))] \\
&= \mathbb{E}[\gamma_1 Y_2 (Z_2 - \mathbb{E}(Z_2))] && \text{since } Z_2 \perp\!\!\!\perp U_1 \\
&= \mathbb{E}\left[\gamma_1 \left(\frac{U_2 + \gamma_2 U_1}{1-\gamma_1\gamma_2} + \frac{\beta_2}{1-\gamma_1\gamma_2} Z_2\right)(Z_2 - \mathbb{E}(Z_2))\right] \\
&= 0 + \mathbb{E}\left[\frac{\gamma_1\beta_2}{1-\gamma_1\gamma_2}\right] \operatorname{var}(Z_2) && \text{since } Z_2 \perp\!\!\!\perp (\beta_2, U, \Gamma)
\end{aligned}$$

and

$$\begin{aligned}\operatorname{cov}(Y_2, Z_2) &= \mathbb{E}\left[\left(\frac{U_2 + \gamma_2 U_1}{1-\gamma_1\gamma_2} + \frac{\beta_2}{1-\gamma_1\gamma_2} Z_2\right)(Z_2 - \mathbb{E}(Z_2))\right] \\ &= 0 + \mathbb{E}\left[\frac{\beta_2}{1-\gamma_1\gamma_2}\right] \operatorname{var}(Z_2) \qquad \text{since } Z_2 \perp\!\!\!\perp (\beta_2, U, \Gamma).\end{aligned}$$

□

Proof of lemma 1. First suppose $Y = \pi' Z$ where $\pi = (A, B)$ and $Z = (Z_0, Z_1, \ldots, Z_K)$ has full support on $\mathbb{R}^{K+1}$. The characteristic function of $Y \mid Z$ is

$$\begin{aligned}\phi_{Y|Z}(t \mid z) &= \mathbb{E}[\exp(itY) \mid Z = z] \\ &= \mathbb{E}[\exp(it(\pi' Z)) \mid Z = z] \\ &= \mathbb{E}[\exp(i(tz)'\pi)] \\ &= \phi_\pi(tz) \\ &= \phi_\pi(tz_0, tz_1, \ldots, tz_K),\end{aligned}$$

where the third line follows since $Z \perp\!\!\!\perp (A, B)$. Thus

$$\phi_\pi(tz) = \phi_{Y|Z}(t \mid z) \quad \text{all } t \in \mathbb{R},\ z \in \operatorname{supp}(Z) = \mathbb{R}^{K+1}.$$

So ϕ_π is completely known and hence the distribution of π is known. For example, setting $t = 1$ shows that we can obtain the entire characteristic function ϕ_π by varying z. Notice that we do not need to vary t at all. Now return to the original problem, $Y = A + B'Z$. This is the same problem we just considered, except that $Z_0 \equiv 1$. Thus we have

$$\phi_\pi(t, tz_1, \ldots, tz_K) = \phi_{Y|Z}(t \mid z) \quad \text{all } t \in \mathbb{R},\ z \in \mathbb{R}^K.$$

In this case, the entire characteristic function ϕ_π is still observed. Suppose we want to learn $\phi_\pi(s_0, \ldots, s_K)$, the characteristic function evaluated at some point $(s_0, \ldots, s_K) \in \mathbb{R}^{K+1}$. If $s_0 \neq 0$, let $t = s_0$ and $z_k = s_k/s_0$. If $s_0 = 0$, then consider a sequence $(t_n, z_{1n}, \ldots, z_{Kn})$ where $t_n \neq 0$, $t_n \to 0$ as $n \to \infty$, and $z_{kn} = s_k/t_n$. Then

$$\begin{aligned}\lim_{n\to\infty} \phi_{Y|Z}(t_n, t_n z_{1n}, \ldots, t_n z_{Kn}) &= \lim_{n\to\infty} \phi_{Y|Z}(t_n, s_1, \ldots, s_K) \\ &= \lim_{n\to\infty} \phi_\pi(t_n, s_1, \ldots, s_K) \\ &= \phi_\pi\left(\lim_{n\to\infty} t_n, s_1, \ldots, s_K\right) \\ &= \phi_\pi(0, s_1, \ldots, s_K),\end{aligned}$$

where the third line follows by continuity of the characteristic function. Thus the distribution of $\pi = (A, B)$ is identified. □

Proof of sufficiency in lemma 2.

1. *Preliminary definitions and notation.* Let L be an arbitrary closed subspace of $\mathbb{R}^{K+1}$. Let $\operatorname{proj}_L : \mathbb{R}^{K+1} \to L$ denote the orthogonal projection of $\mathbb{R}^{K+1}$ onto L. For an arbitrary probability distribution G on $\mathbb{R}^{K+1}$, let G_L denote the *projection* of G onto L, which is

defined as the probability distribution on L such that

$$P_{G_L}(B) \equiv P_G(\text{proj}_L^{-1}(B))$$

for each (measurable) $B \subseteq L$. That is, the probability under G_L of an event B is the probability under G of the event $\text{proj}_L^{-1}(B)$, the set of all elements in $\mathbb{R}^{K+1}$ which project into B.

Let $\ell(\hat{z}) = \{\lambda\hat{z} \in \mathbb{R}^{K+1} : \lambda \in \mathbb{R}\}$ denote the one-dimensional subspace of $\mathbb{R}^{K+1}$ defined by the line passing through the origin and the point $\hat{z} \in \mathbb{R}^{K+1}$. Random coefficient models essentially tell us the projection of the distribution (A, B) onto various lines $\ell(\hat{z})$, and our goal is to recover the original $(K+1)$-dimensional distribution.

2. *Proof.* Let F denote the true distribution of (A, B) and let $\tilde{F}$ denote an observationally equivalent distribution of (A, B). The conditional distribution of $Y \mid Z = z$ is the projection of (A, B) onto the line $\ell(1, z_1, \ldots, z_K)$. Multiplying Y by a scalar λ tells us the projection of (A, B) onto the line $\ell(\lambda, \lambda z_1, \ldots, \lambda z_K)$; alternatively, simply note that $\ell(1, z_1, \ldots, z_K) = \ell(\lambda, \lambda z_1, \ldots, \lambda z_K)$ for any nonzero scalar λ. Thus, since F and $\tilde{F}$ are observationally equivalent, we know that $F_{\ell(\lambda,\lambda z)} = \tilde{F}_{\ell(\lambda,\lambda z)}$ for each $z \in \text{supp}(Z)$ and each $\lambda \in \mathbb{R}$. Let

$$\begin{aligned} R &\equiv \{(\lambda, \lambda z_1, \ldots, \lambda z_K) \in \mathbb{R}^{K+1} : z \in \text{supp}(Z), \lambda \in \mathbb{R}\} \\ &\subseteq \{(\lambda, \lambda z_1, \ldots, \lambda z_K) \in \mathbb{R}^{K+1} : F_{\ell(\lambda,\lambda z)} = \tilde{F}_{\ell(\lambda,\lambda z)}\}. \end{aligned}$$

(Note that these sets are not necessarily equal since $F_{\ell(\lambda,\lambda z)} = \tilde{F}_{\ell(\lambda,\lambda z)}$ might hold for $z \notin \text{supp}(Z)$. Indeed, we shall show that $F = \tilde{F}$, in which case the latter set is strictly larger than the former anytime $\text{supp}(Z) \neq \mathbb{R}^K$.)

For $\hat{z} = (\lambda, \lambda z) \in R$ we have

$$\begin{aligned} \int (\hat{z}'y)^n dF(y) &= \int (t)^n dF_{\ell(\lambda,\lambda z)}(t) \\ &= \int (t)^n d\tilde{F}_{\ell(\lambda,\lambda z)}(t) \\ &= \int (\hat{z}'y)^n d\tilde{F}(y). \end{aligned}$$

These integrals are finite by assumption. The first and third lines follow by a change of variables and the definition of the projection onto a line. The second line follows since $\hat{z} \in R$.

Define the homogeneous polynomial $p_n : \mathbb{R}^{K+1} \to \mathbb{R}$ by

$$p_n(\hat{z}) \equiv \int (\hat{z}'y)^n dF(y) - \int (\hat{z}'y)^n d\tilde{F}(y).$$

Thus we have $p_n(\hat{z}) = 0$ for all $\hat{z} \in R$. That is,

$$R \subseteq S \equiv \{\hat{z} \in \mathbb{R}^{K+1} : p_n(\hat{z}) = 0\}.$$

If p_n is not identically zero then the set S is a hypersurface in $\mathbb{R}^{K+1}$, and thus has Lebesgue measure zero by lemma 3. (Here 'Lebesgue measure' refers to the Lebesgue measure on $\mathbb{R}^{K+1}$.) This implies that R has Lebesgue measure zero. But this is a contradiction: $\text{supp}(Z)$

contains an open ball and thus R contains a cone in $\mathbb{R}^{K+1}$ (see figure 4), which has positive Lebesgue measure.

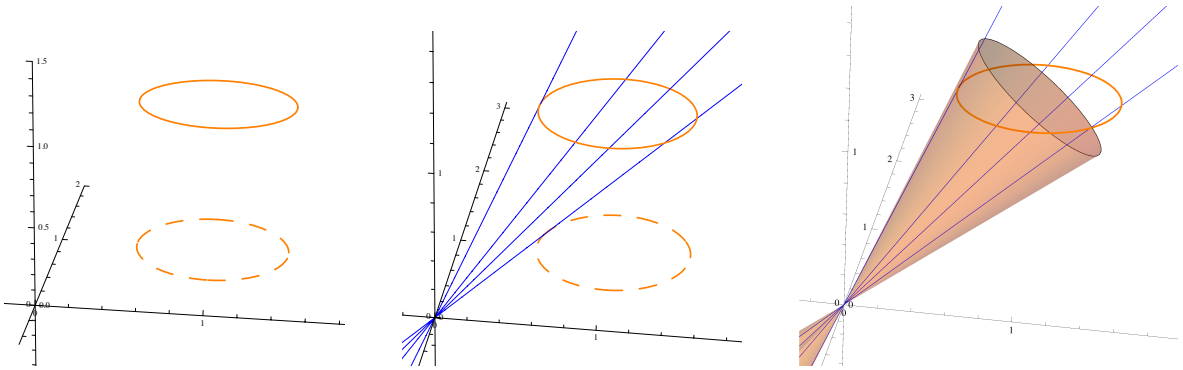


Figure 4: Let $K = 2$. The horizontal plane shows values of (z_1, z_2), while the vertical axis shows 'z_0'. The first plot shows the open ball in $\text{supp}(Z)$ as a dashed circle, which is projected up into the plane $z_0 \equiv 1$, as a solid circle. We know all projections onto lines $\ell(1, z)$ in this set. The second plot shows four example lines, through points near the edge of the set. By scaling all of these points up or down by $\lambda \in \mathbb{R}$, we know all projections onto lines $\ell(\hat{z})$ for points $\hat{z}$ inside an entire cone, as shown in the third plot (the cone drawn is only approximately correct).

Thus p_n must be identically zero. That is,

$$\int (\hat{z}'y)^n dF(y) = \int (\hat{z}'y)^n d\tilde{F}(y)$$

for all $\hat{z} \in \mathbb{R}^{K+1}$ and all natural numbers n. By lemma 4, this implies that F and $\tilde{F}$ have the same moments. Thus $F = \tilde{F}$.

□

Lemma 3. Let $p : \mathbb{R}^K \to \mathbb{R}$ be a polynomial of degree n, not identically zero. Define

$$S = \{z \in \mathbb{R}^K : p(z) = 0\}.$$

Then S has $\mathbb{R}^K$-Lebesgue measure zero.

S is known as a Zariski closed set in Algebraic Geometry, so this lemma states that Zariski closed sets have measure zero.

Proof of lemma 3. This follows from Rossi and Gunning (1965) corollary 10 on page 9. Also see the lemma on page 763 of Okamoto (1973), and Landsberg (2012) page 115. □

Lemma 4. Let F and G be two cdfs on $\mathbb{R}^K$. Then

$$\int (z'y)^n \, dF(y) = \int (z'y)^n \, dG(y) \qquad \text{for all } z \in \mathbb{R}^K, \, n \in \mathbb{N}$$

implies that F and G have the same moments.

This lemma states that knowledge of the moments of the projection onto each line $\ell(z)$ is sufficient for knowledge of the moments of the entire K-dimensional distribution.

Proof of lemma 4. Fix $n \in \mathbb{N}$. Define

$$\begin{aligned} p_F(z) &\equiv \int (z'y)^n \, dF(y) \\ &= \sum_{j_1+\cdots+j_K=n} \binom{n}{j_1 \cdots j_K} z_1^{j_1} \cdots z_K^{j_K} m^F_{j_1,\ldots,j_K}, \end{aligned}$$

where

$$m^F_{j_1,\ldots,j_K} \equiv \int y_1^{j_1} \cdots y_K^{j_K} \, dF(y)$$

are the moments of F. Define $p_G(z)$ likewise. The functions $p_F(z)$ and $p_G(z)$ are polynomials of degree n. By assumption, $p_F = p_G$. Thus the coefficients on the corresponding terms $z_1^{j_1} \cdots z_K^{j_K}$ must be equal:

$$m^F_{j_1,\ldots,j_K} = m^G_{j_1,\ldots,j_K}.$$

This follows by differentiating the identity $p_F(z) \equiv p_G(z)$ in different ways. For example,

$$\frac{\partial^n}{\partial z_1^n} p_F(z) = m^F_{n,0,\ldots,0} = m^G_{n,0,\ldots,0} = \frac{\partial^n}{\partial z_1^n} p_G(z).$$

In general, just apply

$$\frac{\partial^n}{\partial_1^{j_1} \cdots \partial_K^{j_K}} p_F(z) = m^F_{j_1,\ldots,j_K}.$$

n was arbitrary, and thus F and G have the same moments. □

Lemma 5. In lemma 2, assumption (4) implies assumption (3).

Proof of lemma 5. Let P be a probability measure which is uniquely determined by its first n moments. If it is compactly supported (e.g., a bernoulli distribution), then (3) holds immediately; all moments of P actually exist. So suppose it has unbounded support. We prove this case by contrapositive. Suppose P only has its first n moments. Then Tchakaloff's theorem (see theorem 2 in Bayer and Teichmann 2006) implies there is a finitely discretely supported probability distribution Q with the same n moments. (This is perhaps obvious for distributions on $\mathbb{R}$, but the cited theorem shows it holds for probability measures on any $\mathbb{R}^{K+1}$ for any integer $K \geq 1$.) But P is not finitely discretely supported, so $P \neq Q$, and hence (4) does not hold. Thus we have shown that a probability distribution which does not have all its moments cannot be uniquely determined by the set of moments it does have. □

Proof of necessity in lemma 2. By lemma 5, assumption (4) implies assumption (3), and hence it suffices to show that (4) is necessary.

Necessity of assumption (4) for identification of the joint distribution of (A, B) follows by directly applying the counterexample given in theorem 5.4 of Bélisle et al. (1997); see also Cuesta-Albertos et al. (2007) theorem 3.6. The important step in applying theorem 5.4 to random coefficient models is noting that we choose the closed ball K (in their notation) in $\mathbb{R}^{1+\dim(Z)}$ to be outside of the cone passing through $\text{supp}(1, Z)$; e.g. outside of the cone drawn in the third plot of figure 4. Then conclusion (i) of theorem 5.4 shows that the two constructed measures μ and ν have identical projections on all $\dim(Z)$-dimensional subspaces not which do not intersect K. These subspaces include the cone passing through $\text{supp}(1, Z)$. Moreover, having identical projections on a higher dimensional subspace implies that the projections on lower dimensional subspaces—namely, the

one-dimensional lines—are also identical. Hence these two measures μ and ν are observationally equivalent.

Note that if Z had full support then any choice of K would intersect the support of the cone passing through $\{(1, z) \in \mathbb{R}^{1+\dim(Z)} : z \in \text{supp}(Z)\}$. But the theorem only guarantees that the two measures μ and ν have identical projections outside of K; it allows them to have different projections inside K, and hence they will not be observationally identical. This is where theorem 5.4 fails to apply in the full support case.

To see that (4) is also necessary for identification of the marginal distributions, it suffices to choose K slightly more carefully. The basic idea is that in the counterexample, the region K is where we allow our measures to differ. After all, the two measures are not going to be the same, so they have to differ somewhere. In the next step I show that we can choose K to ensure that the measures differ in their projection along one of the axes; this projection is just the marginal distribution of the random coefficient corresponding to that axis.

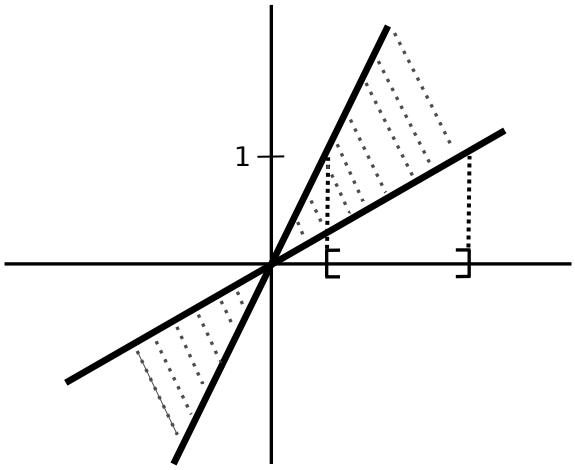


Figure 5: For a scalar regressor, $\dim(Z) = 1$, (A, B) is two-dimensional with support contained in the plane $\mathbb{R}^2$. We observe projections of this bivariate distribution along lines contained in the set R, plotted here as the shaded area. This set is determined by the support of Z, shown as the bracketed interval.

To see this formally, I show how to modify Bélisle et al. (1997)'s proof of theorem 5.4 to obtain the desired result. I use their notation here. Choose K such that it overlaps with one of the axes $2, \ldots, d$, say the kth axis. In the present context of the random coefficient model, this is possible because Z having bounded support implies that the set

$$R \equiv \{(\lambda, \lambda z_1, \ldots, \lambda z_{\dim(Z)}) \in \mathbb{R}^{1+\dim(Z)} : z \in \text{supp}(Z), \lambda \in \mathbb{R}\}$$

(defined as in the proof of sufficiency of lemma 2) intersects the axes $2, \ldots, d$ only at the origin. For example, consider the case where Z is a scalar. Figure 5 plots the set R, where the support of Z is shown as the bracketed interval on the horizontal axis. This figure is similar to figure 4, except here Z is a scalar instead of a 2-vector. The important point here is that because $\text{supp}(Z)$ is bounded above and below, the cone R never intersects the horizontal axis. Hence there always exists a ball K containing the axis but not intersecting R.

Next, choose p (at the beginning of the proof of theorem 5.4, Bélisle et al. (1997) page 783) to lie exactly on this axis. Then, for the function f defined on page 782, $f(p) > 0$. Moreover the point $p + p$ still lies on the axis (since the kth component of p is zero and zero plus zero is still zero).

From the proof of their lemma 5.5 we're working with a function σ whose Fourier transform $\widehat{\sigma}$ is

$$\widehat{\sigma}(t) = \frac{1}{2}[(f * f)(-t) + (f * f)(t)]$$

where $*$ denotes convolution. Next, since $f(p) > 0$, $f \geq 0$, and f is infinitely differentiable, $(f * f)(p + p) > 0$. This implies

$$\begin{aligned}\widehat{\sigma}(p+p) &= \frac{1}{2}[(f * f)(-[p+p]) + (f * f)(p+p)] \\ &> \frac{1}{2}(f * f)(-[p+p]) + 0 \\ &\geq 0\end{aligned}$$

where last line follows since $f \geq 0$.

Thus $p + p$ is in the support of $\widehat{\sigma}$. Importantly, this function σ is defined in Bélisle et al. (1997) such that $\widehat{\sigma} \equiv \widehat{\lambda}_1 - \widehat{\lambda}_2$. Hence $\widehat{\lambda}_1(p + p) \neq \widehat{\lambda}_2(p + p)$. λ_1 and λ_2 are essentially the measures μ and ν we are constructing as our counterexamples (the only difference is that λ_1 and λ_2 are not normalized to have measure one). Hence $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are essentially just the characteristic functions of the two measures μ and ν. So $\widehat{\lambda}_1(p + p) \neq \widehat{\lambda}_2(p + p)$ implies that these characteristic functions are *different* for projections passing through $p + p$. That is, their projections onto this axis are different. Hence they have different marginal distributions of Z_k.

Finally, consider the intercept A. If $0 \in \text{supp}(Z)$ then the distribution of A is point identified from the distribution of $Y \mid Z = 0$. In this case we can also see how the above nonidentification proof would no longer apply. For example, consider figure 5. If $0 \in \text{supp}(Z)$, then the cone would cover the vertical axis, which would prevent us from choosing a K that overlaps with the the vertical axis. On the other hand, if $0 \notin \text{supp}(Z)$, then the above proof applies equally to the 1st axis (corresponding to the intercept), thus showing that the marginal distribution of A is not point identified in this case. □

Proof of theorem 1. The Beran and Millar (1994) proof relied on the assumption that the random coefficients had compact support, which implies that their characteristic function is analytic. Assumptions (3) and (4) are not sufficient for the characteristic function to be analytic, and hence their proof by analytic continuation does not apply. I instead use the proof strategy from lemma 2. For simplicity I consider the case $K_1 = K_2 = 1$, where there is only one covariate per equation. The multivariate case only requires additional notation. For any $(t_1, t_2) \in \mathbb{R}^2$, consider the linear combination

$$t_1 Y_1 + t_2 Y_2 = t_1 A_1 + t_2 A_2 + t_1 Z_1 B_1 + t_2 Z_2 B_2.$$

If we consider the distribution of this linear combination conditional on $(Z_1, Z_2) = (z_1, z_2)$, we see that we are observing the distribution of the linear combination

$$t_1 A_1 + t_2 A_2 + t_1 z_1 B_1 + t_2 z_2 B_2.$$

Put differently, the characteristic function of $(Y_1, Y_2) \mid (Z_1, Z_2)$ is

$$\begin{aligned}
\phi_{Y_1,Y_2|Z_1,Z_2}(t_1, t_2 \mid z_1, z_2) &= \mathbb{E}[\exp(i[t_1 Y_1 + t_2 Y_2]) \mid Z_1 = z_1, Z_2 = z_2] \\
&= \mathbb{E}[\exp(i[t_1 A_1 + t_1 Z_1 B_1 + t_2 A_2 + t_2 Z_2 B_2]) \mid Z_1 = z_1, Z_2 = z_2] \\
&= \mathbb{E}[\exp(i[t_1 A_1 + t_2 A_2 + t_1 z_1 B_1 + t_2 z_2 B_2]) \mid Z_1 = z_1, Z_2 = z_2] \\
&= \mathbb{E}[\exp(i[t_1 A_1 + t_2 A_2 + t_1 z_1 B_1 + t_2 z_2 B_2])] \\
&= \phi_{A_1,A_2,B_1,B_1}(t_1, t_2, t_1 z_1, t_2 z_2).
\end{aligned}$$

Define

$$R \equiv \{(t_1, t_2, t_1 z_1, t_2 z_2) \in \mathbb{R}^4 : (z_1, z_2) \in \text{supp}(Z_1, Z_2), t_1, t_2 \in \mathbb{R}\}.$$

Let F and $\tilde{F}$ denote observationally equivalent distributions of (A, B). Then

$$R \subseteq \{(t_1, t_2, t_1 z_1, t_2 z_2) \in \mathbb{R}^4 : F_{\ell(t_1,t_2,t_1 z_1,t_2 z_2)} = \tilde{F}_{\ell(t_1,t_2,t_1 z_1,t_2 z_2)}\}$$

where $\ell(\cdot)$ denotes a line in $\mathbb{R}^4$ and $F_{\ell(\cdot)}$ the projection onto that line, both as defined in the proof of lemma 2. The proof now continues exactly as in the proof of lemma 2. It concludes by noting that R does not have Lebesgue measure zero because $\text{supp}(Z_1, Z_2)$ contains an open ball in $\mathbb{R}^2$, and thus R contains an open ball in $\mathbb{R}^4$. That concludes the proof of sufficiency of assumptions (1)–(4).

The proof of necessity of the moment conditions follows because the SUR model nests the single equation model.

Finally, to see that functional relationships between components of (Z_1, Z_2) result in a lack of point identification, I apply a counterexample from Cuesta-Albertos et al. (2007). Without loss of generality it suffices to consider the case $Z_1 \equiv Z_2$; if the functional relationship is not the identity then we can simply redefine our covariates to make it so. Likewise, it suffices to consider the case where there is one covariate per equation, because the multivariate model nests the single-variate model. Hence we consider the model

$$\begin{aligned}
Y_1 &= A_1 + B_1 Z \\
Y_2 &= A_2 + B_2 Z,
\end{aligned}$$

where $Z \equiv Z_1 \equiv Z_2$. By a similar argument as above, we have

$$\phi_{Y_1,Y_2|Z}(t_1, t_2 \mid z) = \phi_{A_1,A_2,B_1,B_2}(t_1, t_2, t_1 z, t_2 z)$$

for any $t_1, t_2 \in \mathbb{R}$ and $z \in \text{supp}(Z)$. For simplicity assume $\text{supp}(Z) = \mathbb{R}$; the lack of point identification result holds even in this case. Thus we see that the characteristic function of (A_1, A_2, B_1, B_2) is known on the set

$$R \equiv \{(t_1, t_2, t_1 z, t_2 z) \in \mathbb{R}^4 : t_1, t_2 \in \mathbb{R}, z \in \text{supp}(Z)\}.$$

Define the homogeneous polynomial $p : \mathbb{R}^4 \to \mathbb{R}$ by

$$p(x) = x_1 x_4 - x_2 x_3.$$

Then

$$R \subseteq \{x \in \mathbb{R}^4 : p(x) = 0\}.$$

To see this, let $x \in R$. Then there exists $t_1, t_2 \in \mathbb{R}$ and $z \in \text{supp}(Z)$ such that $(x_1, x_2, x_3, x_4) =$

$(t_1, t_2, t_1 z, t_2 z)$. Hence

$$\begin{aligned} p(x) &= x_1 x_4 - x_2 x_3 \\ &= t_1 t_2 z - t_2 t_1 z \\ &= 0. \end{aligned}$$

So $x \in \{x \in \mathbb{R}^4 : p(x) = 0\}$. Note that p is not identically zero. Thus R has $\mathbb{R}^4$-Lebesgue measure zero by lemma 3. That is, the characteristic function of (A_1, A_2, B_1, B_2) is point identified only on a set of measure zero. This is the key problem. The counterexample given in theorem 3.5 of Cuesta-Albertos et al. (2007) shows that knowledge of a characteristic function on such sets (specifically, projective hypersurfaces) of measure zero is not sufficient to pin down the underlying distribution. Indeed, they show that this is true even if we assumed the underlying distribution has compact support. The lack of point identification of the joint distribution of (A_1, A_2, B_1, B_2) follows. □

Proof of theorem 2. The proof has three steps: (1) Identify the joint distribution of linear combinations of the reduced form coefficients, (2) Identify the marginal distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$, and (3) Show that A5 is necessary when $\text{supp}(Z \mid X = x)$ is bounded.

1. Fix an $x \in \text{supp}(X)$. For any $z \in \text{supp}(Z \mid X = x)$, we observe the joint distribution of (Y_1, Y_2) given $Z = z, X = x$, which is given by the reduced form system

$$\begin{aligned} Y_1 &= \frac{U_1 + \gamma_1 U_2 + (\delta_1 + \gamma_1 \delta_2)' x}{1 - \gamma_1 \gamma_2} + \frac{\beta_1}{1 - \gamma_1 \gamma_2} z_1 + \frac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2} z_2 \\ Y_2 &= \frac{U_2 + \gamma_2 U_1 + (\delta_2 + \gamma_2 \delta_1)' x}{1 - \gamma_1 \gamma_2} + \frac{\gamma_2 \beta_1}{1 - \gamma_1 \gamma_2} z_1 + \frac{\beta_2}{1 - \gamma_1 \gamma_2} z_2. \end{aligned}$$

   Define

$$\begin{aligned} \pi_1 \equiv (\pi_{11}, \pi_{12}, \pi_{13}) &\equiv \left(\frac{U_1 + \gamma_1 U_2 + (\delta_1 + \gamma_1 \delta_2)' x}{1 - \gamma_1 \gamma_2}, \frac{\beta_1}{1 - \gamma_1 \gamma_2}, \frac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2} \right) \\ \pi_2 \equiv (\pi_{21}, \pi_{22}, \pi_{23}) &\equiv \left(\frac{U_2 + \gamma_2 U_1 + (\delta_2 + \gamma_2 \delta_1)' x}{1 - \gamma_1 \gamma_2}, \frac{\gamma_2 \beta_1}{1 - \gamma_1 \gamma_2}, \frac{\beta_2}{1 - \gamma_1 \gamma_2} \right). \end{aligned}$$

   For $(t_1, t_2) \in \mathbb{R}^2$, we have

$$t_1 Y_1 + t_2 Y_2 = (t_1 \pi_{11} + t_2 \pi_{21}) + (t_1 \pi_{12} + t_2 \pi_{22}) z_1 + (t_1 \pi_{13} + t_2 \pi_{23}) z_2.$$

   By A3, A4, and A5, we can apply lemma 2 to show that, for any $z_1 \in \text{supp}(Z_1 \mid X = x)$, the joint distribution of

$$([t_1 \pi_{11} + t_2 \pi_{21}] + [t_1 \pi_{12} + t_2 \pi_{22}] z_1, \quad t_1 \pi_{13} + t_2 \pi_{23})$$

   given $X = x$ is identified, for each $(t_1, t_2) \in \mathbb{R}^2$. Hence the distribution of $t_1 \pi_{13} + t_2 \pi_{23}$ is identified for each $(t_1, t_2) \in \mathbb{R}^2$. Likewise, the joint distribution of

$$([t_1 \pi_{11} + t_2 \pi_{21}] + [t_1 \pi_{13} + t_2 \pi_{23}] z_2, \quad t_1 \pi_{12} + t_2 \pi_{22})$$

   given $X = x$ is identified, for each $(t_1, t_2) \in \mathbb{R}^2$. Hence the distribution of $t_1 \pi_{12} + t_2 \pi_{22}$ is identified for each $(t_1, t_2) \in \mathbb{R}^2$.

2. Consider the term $t_1\pi_{13} + t_2\pi_{23}$. The distribution of this scalar random variable is identified for each $(t_1, t_2) \in \mathbb{R}^2$, given $X = x$. By definition, the characteristic function of (π_{13}, π_{23}) is

$$\phi_{\pi_{13},\pi_{23}}(t_1, t_2) = \mathbb{E}[\exp(i(t_1\pi_{13} + t_2\pi_{23}))].$$

The right hand side is identified for each $(t_1, t_2) \in \mathbb{R}^2$ and hence the characteristic function $\phi_{\pi_{13},\pi_{23}}$ is identified. Thus the joint distribution of (π_{13}, π_{23}) is identified, given $X = x$. Likewise, the joint distribution of (π_{12}, π_{22}) is identified, given $X = x$.

Since the joint distribution of

$$(\pi_{13}, \pi_{23}) = \left(\frac{\beta_2}{1 - \gamma_1\gamma_2}\gamma_1, \frac{\beta_2}{1 - \gamma_1\gamma_2}\right)$$

is identified, given X, lemma 6 implies that $\gamma_1 \mid X$ is identified.[6] Likewise, since the joint distribution of

$$(\pi_{12}, \pi_{22}) = \left(\frac{\beta_1}{1 - \gamma_1\gamma_2}, \frac{\beta_1}{1 - \gamma_1\gamma_2}\gamma_2\right)$$

is identified, given X, lemma 6 implies that $\gamma_2 \mid X$ is identified.

3. Consider the following special case of system (1):

$$\begin{aligned} Y_1 &= \gamma_1 Y_2 + U_1 \\ Y_2 &= \beta_2 Z_2 \end{aligned}$$

where $\delta_1, \delta_2, \gamma_2, \beta_1, U_2$ are all identically zero. Suppose β_2 is a constant. Then this model is really just a single equation model with exogeneity:

$$Y_1 = \gamma_1\beta_2 Z_2 + U_1$$

where β_2 is a known constant. $\text{supp}(Z \mid X = x)$ bounded implies that $\text{supp}(Z_2 \mid X = x)$ is bounded. Suppose that A5 does not hold. Then the distribution of (γ_1, U_1) is not uniquely determined by its moments. Hence the proof of lemma 2 shows that we can construct two observationally equivalent distributions of (γ_1, U_1) which have distinct marginal distributions of γ_1.

□

Lemma 6. Let Y and X be random variables. Assume X does not have a mass point at zero. Suppose the joint distribution of (YX, X) is observed. Then the joint distribution of (Y, X) is identified, and hence the distribution of Y is identified.

Proof of lemma 6. The distribution of X is identified directly from the observed marginal distri-

[6] Alternatively, note that $\gamma_1 = \pi_{13}/\pi_{23}$. The distribution of the right hand side random variable is identified, and thus γ_1 is identified. Lemma 6 simply makes this argument more formal by showing how to write the cdf of γ_1 directly in terms of observed cdfs. A similar argument applies to $\gamma_2 = \pi_{22}/\pi_{12}$.

bution of (YX, X). Next, we have

$$\begin{aligned}\mathbb{P}(YX \leq yx \mid X = x) &= \mathbb{P}(Yx \leq yx \mid X = x) \\ &= \begin{cases} \mathbb{P}(Y \leq y \mid X = x) & \text{if } x > 0 \\ 1 & \text{if } x = 0 \\ \mathbb{P}(Y \geq y \mid X = x) & \text{if } x < 0. \end{cases}\end{aligned}$$

Thus, for $x > 0$,

$$\mathbb{P}(Y \leq y \mid X = x) = \mathbb{P}(YX \leq yx \mid X = x)$$

and, for $x < 0$,

$$\mathbb{P}(Y \leq y \mid X = x) = 1 - \mathbb{P}(YX \leq yx \mid X = x) + \mathbb{P}(YX = yx \mid X = x).$$

So $F_{Y|X}(y \mid x) = \mathbb{P}(Y \leq y \mid X = x)$ is identified for all $x \neq 0$. Consequently, for $t > 0$,

$$\begin{aligned}F_{Y,X}(y, t) &= \mathbb{P}(Y \leq y, X \leq t) \\ &= \int_{-\infty}^{t} F_{Y|X}(y \mid x) dF_X(x) \\ &= \int_{\{t>x>0\}} F_{Y|X}(y \mid x) dF_X(x) + \int_{\{x<0\}} F_{Y|X}(y \mid x) dF_X(x) + \int_{\{x=0\}} F_{Y|X}(y \mid x) dF_X(x) \\ &= \int_{\{t>x>0\}} F_{Y|X}(y \mid x) dF_X(x) + \int_{\{x<0\}} F_{Y|X}(y \mid x) dF_X(x),\end{aligned}$$

where the second line follows by iterated expectations and the fourth line follows since X does not have a mass point at zero. The last line is identified. The result is analogous for $t \leq 0$. Hence $F_{Y,X}$ is identified. □

Proof of proposition 1. I suppress conditioning on X everywhere. Here we show that A6 implies A5.2, (π_1, π_2) is uniquely determined by its moments. As discussed earlier, this is also sufficient for the existence of all absolute moments. Petersen (1982, theorem 3, page 363) showed that, for an arbitrary random vector Y, if the coordinate random variables Y_j are uniquely determined by their moments, then Y is uniquely determined by its moments. Thus it suffices to show that each of the components of (π_1, π_2) are separately uniquely determined by their moments. I will only consider the three components π_{11}, π_{12}, π_{13}; the proof for the components of π_2 is analogous.

The moment generating function of π_{12} is, for small enough $t > 0$,

$$\begin{aligned}\text{MGF}_{\pi_{12}}(t) &= \mathbb{E}[\exp(t\pi_{12})] \\ &= \mathbb{E}[\exp(t\beta_1/(1 - \gamma_1\gamma_2))] \\ &= \int_{\beta_1 \geq 0} \exp\left(t\beta_1 \frac{1}{1 - \gamma_1\gamma_2}\right) \, dF_{\beta_1,\gamma_1,\gamma_2} + \int_{\beta_1 < 0} \exp\left(t\beta_1 \frac{1}{1 - \gamma_1\gamma_2}\right) \, dF_{\beta_1,\gamma_1,\gamma_2} \\ &\leq \int_{\beta_1 \geq 0} \exp\left([t/\tau]\beta_1\right) \, dF_{\beta_1,\gamma_1,\gamma_2} + \int_{\beta_1 < 0} \exp\left([-t/\tau]\beta_1\right) \, dF_{\beta_1,\gamma_1,\gamma_2} \\ &\leq \text{MGF}_{\beta_1}(-t/\tau) + \text{MGF}_{\beta_1}(t/\tau) \\ &< \infty\end{aligned}$$

where the fourth line follows by A6.1 and the last line since the MGF of β_1 exists by A6.2 and lemma

8. An analogous argument holds for small enough $t < 0$. Thus the moment generating function of π_{12} exists in a neighborhood of zero and hence π_{12} is uniquely determined by its moments. An analogous argument shows that the moment generating function of π_{13} exists in a neighborhood of zero, since the MGF of $\beta_1\gamma_2 \mid X$ exists in a neighborhood of zero by A6.2 and lemma 8.

Finally, consider the moment generating function of π_{11}:

$$\begin{aligned}\mathrm{MGF}_{\pi_{11}}(t) &= \mathbb{E}[\exp(t\pi_{11})] \\ &= \mathbb{E}\left[\exp\left(t\left[\frac{1}{1-\gamma_1\gamma_2}U_1 + \frac{1}{1-\gamma_1\gamma_2}(\gamma_1 U_2) + \frac{1}{1-\gamma_1\gamma_2}\delta_1' x + \frac{1}{1-\gamma_1\gamma_2}(\gamma_1\delta_2)' x\right]\right)\right].\end{aligned}$$

A similar argument to above splits the support of the random coefficients into $2^4 = 16$ pieces, one for each combination of signs of the four terms $U_1, \gamma_1 U_2, \delta_1' x, (\gamma_1\delta_2)' x$, and then uses A6.1 to eliminate the denominator term. That leaves us with a sum of the moment generating function of $(U_1, \gamma_1 U_2, \delta_1, \gamma_1\delta_2)$ evaluated at various points. For small enough t, each of these MGFs exists by assumption A6.2 and lemma 8. Thus the moment generating function of π_{11} exists in a neighborhood of zero and hence π_{11} is uniquely determined by its moments. □

Proof of proposition 2. Suppress conditioning on X.

1. Suppose assumption A6.2′ holds. Then it follows immediately from lemmas 7 and 8 below that A6.2 holds.

2. Next suppose assumption A6.2″ holds. That A6.2 holds follows by a proof similar to that of proposition 1 above. For $t > 0$, the MGF

$$\mathrm{MGF}_{\beta_2\gamma_1}(t) = \mathbb{E}[\exp(t\beta_2\gamma_1)]$$

can be written as a sum of two pieces depending on the sign of β_2, at which point γ_1 can be replaced by either M or $-M$. The result then follows since the MGF of β_2 exists in a neighborhood of zero. Likewise for $\beta_1\gamma_2$, $(U_1, \delta_1, \gamma_1 U_2, \gamma_1\delta_2)$, and $(U_2, \delta_2, \gamma_2 U_1, \gamma_2\delta_1)$.

□

Lemma 7. Let X and Y be random variables with sub-Gaussian tails. Then XY has subexponential tails.

Proof of lemma 7. Let $t > 0$. We have

$$\begin{aligned}\mathbb{P}(|XY| > t) &\leq \mathbb{P}(|X| > \sqrt{t}) + \mathbb{P}(|Y| > \sqrt{t}) \\ &\leq C_x \exp[-c_x(\sqrt{t})^2] + C_y \exp[-c_y(\sqrt{t})^2] \\ &= C_x \exp(-c_x t) + C_y \exp(-c_y t) \\ &\leq (C_x + C_y)\exp(-\min\{c_x, c_y\}t) \\ &\equiv C\exp(-ct).\end{aligned}$$

□

Lemma 8. Let $X_1, \ldots, X_n$ be random variables with subexponential tails. Then the moment generating function of $(X_1, \ldots, X_n)$ exists in a neighborhood of zero.

Proof of lemma 8. This result is related to Peterson's (1982) result that if the components of a random vector are uniquely determined by their moments then the vector itself is uniquely determined by its moments. The MGF existing in a neighborhood of zero implies that the distribution is uniquely determined by its moments, but the converse does not hold. Hence the current lemma is not exactly the same as Peterson's result, because it makes a stronger assumption, but obtains a stronger conclusion.

We already know that the result is true for a single random variable; $n = 1$ (e.g., this can be shown using the same idea as the following). Hence the purpose of this lemma is to show that it is also true for a vector of random variables. It suffices to show the result holds for just two random variables X and Y; the general case extends immediately.

Let $t_1, t_2 \in \mathbb{R}$ be nonzero. The MGF of (X, Y) is

$$\begin{aligned}
\mathrm{MGF}_{X,Y}(t_1, t_2) &= \mathbb{E}[\exp(t_1 X + t_2 Y)] \\
&= \int_0^\infty \mathbb{P}[\exp(t_1 X + t_2 Y) > s]\, ds \\
&= \int_0^1 \mathbb{P}[\exp(t_1 X + t_2 Y) > s]\, ds + \int_1^\infty \mathbb{P}[\exp(t_1 X + t_2 Y) > s]\, ds \\
&\leq 1 + \int_1^\infty \mathbb{P}[\exp(t_1 X + t_2 Y) > s]\, ds \\
&= 1 + \int_1^\infty \mathbb{P}[t_1 X + t_2 Y > \log(s)]\, ds.
\end{aligned}$$

The second line follows since $\exp(t_1 X + t_2 Y)$ is a nonnegative random variable. Next we note that any linear combination of X and Y also has subexponential tails:

$$\begin{aligned}
\mathbb{P}(|t_1 X + t_2 Y| > s) &\leq \mathbb{P}(|t_1 X| > s/2) + \mathbb{P}(|t_2 Y| > s/2) \\
&= \mathbb{P}\left(|X| > \frac{s}{2|t_1|}\right) + \mathbb{P}\left(|Y| > \frac{s}{2|t_2|}\right) \\
&\leq C_x \exp\left(-c_x \frac{s}{2|t_1|}\right) + C_y \exp\left(-c_y \frac{s}{2|t_2|}\right) \\
&\leq (C_x + C_y) \exp\left(-\min\left\{\frac{c_x}{2|t_1|}, \frac{c_y}{2|t_2|}\right\} s\right).
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{MGF}_{X,Y}(t_1, t_2) &\leq 1 + C \int_1^\infty \exp\left(-\min\left\{\frac{c_x}{2|t_1|}, \frac{c_y}{2|t_2|}\right\} \log(s)\right) ds \\
&= 1 + C \int_1^\infty s^{-\min\left\{\frac{c_x}{2|t_1|}, \frac{c_y}{2|t_2|}\right\}}\, ds.
\end{aligned}$$

If t_1 and t_2 are both very small, then the exponent

$$\min\left\{\frac{c_x}{2|t_1|}, \frac{c_y}{2|t_2|}\right\}$$

will be very large, and hence the integral will be finite, because

$$\int_1^\infty \frac{1}{x^p}\, dx < \infty$$

for any $p > 1$. Thus $\text{MGF}_{X,Y}(t_1, t_2)$ exists in an $\mathbb{R}^2$-neighborhood of $(0, 0)$. □

Derivations regarding stability of the equilibrium. Let

$$C = BZ + DX + U.$$

Let Y denote the equilibrium value, $Y = \Gamma Y + C$. Then

$$\begin{aligned} Y_t &= \Gamma Y_{t-1} + C \\ &= \Gamma Y_{t-1} + Y - \Gamma Y \end{aligned}$$

which implies

$$(Y_t - Y) = \Gamma(Y_{t-1} - Y)$$

or $\tilde{Y}_t = \Gamma \tilde{Y}_{t-1}$ where $\tilde{Y}_t = Y_t - Y$ is the deviation from equilibrium. The characterization of global stability now follows immediately from the fact that $\tilde{Y}_t \to 0$ if and only if all eigenvalues of Γ have moduli smaller than 1, which is part (ii) of theorem 4.13 on page 187 of Elaydi (2005). In the present two equation system, we can go further and obtain the explicit characterization that global stability holds if and only if $|\gamma_1 \gamma_2| < 1$ by applying equation 4.3.9 on page 188 of Elaydi (2005). □

Proof of theorem 3. The proof strategy follows the same two steps as in the proof of theorem 2.

1. Use lemma 1 instead of lemma 2 to identify the joint distribution of

   $$(t_1\pi_{11} + t_2\pi_{21}, t_1\pi_{12} + t_2\pi_{22}, t_1\pi_{13} + t_2\pi_{23})$$

   given $X = x$. This step uses A3 and A4′.

2. As in theorem 3.

□

Proof of proposition 3. Throughout the proof we condition all statements on $X = x$ for some $x \in \text{supp}(X)$. There are four steps to the proof: (1) Recall the results on identification of the distribution of reduced form coefficients from the proof of theorems 2 and 3, (2) show that the ratio β_1/β_2 is identified, (3) show that the joint distribution of $(\gamma_1, \gamma_2) \mid X = x$ is identified, and finally (4) show that (β_1, β_2) are identified.

1. From the proof of either theorem 1 or 2, we know that the joint distribution of the reduced form coefficients

   $$(t_1\pi_{11} + t_2\pi_{21}, t_1\pi_{12} + t_2\pi_{22}, t_1\pi_{13} + t_2\pi_{23})$$

   given $X = x$ is identified, for each $(t_1, t_2) \in \mathbb{R}^2$, where we used that $\text{supp}(Z_1, Z_2 \mid X)$ contains an open ball in $\mathbb{R}^2$. In particular, this implies that the marginal distributions of π_{12} and of π_{13} given $X = x$ are identified.

2. Next I show that the scale β_1/β_2 is identified. This would be immediate if the joint distribution of (π_{12}, π_{13}) was known at this step, but it is not. Instead, observe that if $\text{sign}(\beta_1/\beta_2) > 0$

then

$$\begin{aligned} F_{\pi_{12}}\left(t\frac{\beta_1}{\beta_2}\right) &= \mathbb{P}\left(\frac{\beta_1}{1-\gamma_1\gamma_2} \leq t\frac{\beta_1}{\beta_2}\right) \\ &= \mathbb{P}\left(\frac{\beta_2}{\beta_1}\frac{\beta_1}{1-\gamma_1\gamma_2} \leq t\right) \\ &= \mathbb{P}\left(\frac{\beta_2}{1-\gamma_1\gamma_2} \leq t\right) \\ &= F_{\pi_{13}}(t) \quad \text{all } t \in \mathbb{R}, \end{aligned}$$

whereas if $\text{sign}(\beta_1/\beta_2) < 0$ then

$$\begin{aligned} F_{\pi_{12}}\left(t\frac{\beta_1}{\beta_2}\right) &= \mathbb{P}\left(\frac{\beta_1}{1-\gamma_1\gamma_2} \leq t\frac{\beta_1}{\beta_2}\right) \\ &= \mathbb{P}\left(\frac{\beta_2}{\beta_1}\frac{\beta_1}{1-\gamma_1\gamma_2} \geq t\right) \\ &= \mathbb{P}\left(\frac{\beta_2}{1-\gamma_1\gamma_2} \geq t\right) \\ &= 1 - F_{\pi_{13}}(t) + \mathbb{P}(\pi_{13} = t) \quad \text{all } t \in \mathbb{R}. \end{aligned}$$

Suppose that the sign of β_1/β_2 is identified. I will show that this implies that β_1/β_2 itself is identified. First suppose $\text{sign}(\beta_1/\beta_2) > 0$. Then, by the calculations above,

$$F_{\pi_{12}}\left(t\frac{\beta_1}{\beta_2}\right) = F_{\pi_{13}}(t) \quad \text{all } t \in \mathbb{R}.$$

Let $r \in \mathbb{R}$ be such that

$$F_{\pi_{12}}(tr) = F_{\pi_{13}}(t) \quad \text{all } t \in \mathbb{R}.$$

Such an r exists, since $r = \beta_1/\beta_2$ satisfies the above equation. I will show that r is unique, and hence $r = \beta_1/\beta_2$ is identified. Suppose by way of contradiction that there is some $\tilde{r} \neq r$ with

$$F_{\pi_{12}}(t\tilde{r}) = F_{\pi_{13}}(t) \quad \text{all } t \in \mathbb{R}.$$

Suppose without loss of generality that $\tilde{r} > r$. Then

$$F_{\pi_{12}}(tr) = F_{\pi_{12}}(t\tilde{r}) \quad \text{all } t \in \mathbb{R}.$$

If π_{12} has some continuous variation, then there is some point $\bar{t} \neq 0$, so that $F_{\pi_{12}}$ is invertible in a neighborhood of $\bar{t}$. By inverting $F_{\pi_{12}}$ around that $\bar{t}$, we must have $r = \tilde{r}$, a contradiction. If π_{12} has no continuous variation, then π_{12} is discretely distributed. Let s denote a support point. Let $\bar{t} = s/\tilde{r}$. Then

$$\begin{aligned} F_{\pi_{12}}(\bar{t}\tilde{r}) &= \mathbb{P}(\pi_{12} \leq s) \\ &> \mathbb{P}\left(\pi_{12} \leq s\frac{r}{\tilde{r}}\right) \\ &= F_{\pi_{12}}(\bar{t}r) \end{aligned}$$

where the second line follows since $r/\tilde{r} < 1$ and s is a support point of the discretely distributed

π_{12}. This is a contradiction to $F_{\pi_{12}}(\bar{t}\tilde{r}) = F_{\pi_{12}}(\bar{t}r)$ for all $\bar{t} \in \mathbb{R}$.

Next suppose that $\text{sign}(\beta_1/\beta_2) < 0$, so that

$$F_{\pi_{12}}\left(t\frac{\beta_1}{\beta_2}\right) = 1 - F_{\pi_{13}}(t) + \mathbb{P}(\pi_{13} = t) \quad \text{all } t \in \mathbb{R}.$$

Let $r \in \mathbb{R}$ be such that

$$F_{\pi_{12}}(tr) = 1 - F_{\pi_{13}}(t) + \mathbb{P}(\pi_{13} = t) \quad \text{all } t \in \mathbb{R}.$$

Such an r exists since β_1/β_2 satisfies this equation. Let $\tilde{r} \neq r$ also satisfy this equation. Then

$$F_{\pi_{12}}(tr) = F_{\pi_{12}}(t\tilde{r}) \quad \text{all } t \in \mathbb{R}.$$

Now proceed as above. Thus, if the sign of β_1/β_2 is identified, the magnitude of β_1/β_2 is identified.

Next I show that assumption (ii) implies the sign of β_1/β_2 is identified. Note that

$$F_{\pi_{12}}(0) = \mathbb{P}\left(\beta_1\frac{1}{1-\gamma_1\gamma_2} \leq 0\right) = \begin{cases} \mathbb{P}[1/(1-\gamma_1\gamma_2) \leq 0] & \text{if } \beta_1 > 0 \\ 1 - \mathbb{P}[1/(1-\gamma_1\gamma_2) < 0] & \text{if } \beta_1 < 0 \end{cases}$$

and

$$F_{\pi_{23}}(0) = \mathbb{P}\left(\beta_2\frac{1}{1-\gamma_1\gamma_2} \leq 0\right) = \begin{cases} \mathbb{P}[1/(1-\gamma_1\gamma_2) \leq 0] & \text{if } \beta_2 > 0 \\ 1 - \mathbb{P}[1/(1-\gamma_1\gamma_2) < 0] & \text{if } \beta_2 < 0. \end{cases}$$

Thus $\text{sign}(\beta_1/\beta_2) > 0$ implies $F_{\pi_{12}}(0) = F_{\pi_{23}}(0)$. Moreover, $\text{sign}(\beta_1/\beta_2) < 0$ implies $F_{\pi_{12}}(0) \neq F_{\pi_{23}}(0)$. To see this, suppose by way of contradiction that $F_{\pi_{12}}(0) = F_{\pi_{23}}(0)$. Then, since $\text{sign}(\beta_1/\beta_2) < 0$,

$$\mathbb{P}[1/(1-\gamma_1\gamma_2) \leq 0] = 1 - \mathbb{P}[1/(1-\gamma_1\gamma_2) < 0],$$

which is equivalent to

$$\mathbb{P}[1/(1-\gamma_1\gamma_2) \leq 0] + \mathbb{P}[1/(1-\gamma_1\gamma_2) \leq 0] = 1,$$

since the strictly inequality becomes a weak inequality due to $\mathbb{P}[1/(1-\gamma_1\gamma_2) = 0] = 0$, which holds by A1. This, in turn, implies $\mathbb{P}[1/(1-\gamma_1\gamma_2) \leq 0] = 1/2$. But this is a contradiction since

$$\begin{aligned} \mathbb{P}\left(\frac{1}{1-\gamma_1\gamma_1} \leq 0\right) &= \mathbb{P}(1-\gamma_1\gamma_1 \leq 0) \\ &= \mathbb{P}(\gamma_1\gamma_2 \geq 1) \\ &\neq \frac{1}{2} \qquad \text{by assumption (ii).} \end{aligned}$$

Thus, $\text{sign}(\beta_1/\beta_2) > 0$ if and only if $F_{\pi_{12}}(0) = F_{\pi_{23}}(0)$.

3. I thank Daniel Wilhelm for suggesting the following analysis. By step 1, the joint distribution of

$$(t_1\pi_{11} + t_2\pi_{21}, \quad t_1\pi_{12} + t_2\pi_{22}, \quad t_1\pi_{13} + t_2\pi_{23})$$

is identified. Thus we know the joint characteristic function of the second and third components:

$$\phi_{t_1\pi_{12}+t_2\pi_{22},t_1\pi_{13}+t_2\pi_{23}}(s_1,s_2) = \mathbb{E}\Big(\exp\Big(i\big[s_1(t_1\pi_{12}+t_2\pi_{22})+s_2(t_1\pi_{13}+t_2\pi_{23})\big]\Big)\Big).$$

The key step now is to observe that

$$\pi_{23} = \pi_{12}\frac{\beta_2}{\beta_1}$$

and hence

$$\begin{aligned}
\phi_{t_1\pi_{12}+t_2\pi_{22},t_1\pi_{13}+t_2\pi_{23}}(s_1,s_2) &= \mathbb{E}\left(\exp\left(i\left[s_1(t_1\pi_{12}+t_2\pi_{22})+s_2\left(t_1\pi_{13}+t_2\pi_{12}\frac{\beta_2}{\beta_1}\right)\right]\right)\right) \\
&= \mathbb{E}\left(\exp\left(i\left[\left(s_1t_1+s_2t_2\frac{\beta_2}{\beta_1}\right)\pi_{12}+s_1t_2\pi_{22}+s_2t_1\pi_{13}\right]\right)\right) \\
&= \phi_{\pi_{12},\pi_{22},\pi_{13}}\left(s_1t_1+s_2t_2\frac{\beta_2}{\beta_1}, s_1t_2, s_2t_1\right).
\end{aligned}$$

We will show that for a set of $(x_1,x_2,x_3)\in\mathbb{R}^3$ of positive Lebesgue measure, there exists $(s_1,s_2,t_1,t_2)\in\mathbb{R}^4$ such that

$$(x_1,x_2,x_3) = \left(s_1t_1+s_2t_2\frac{\beta_2}{\beta_1}, s_1t_2, s_2t_1\right)$$

Consequently the characteristic function of $(\pi_{12},\pi_{22},\pi_{13})$ is known on a set of positive Lebesgue measure. Hence, by an argument identical to the proof of lemma 2, this shows that the joint distribution of $(\pi_{12},\pi_{22},\pi_{13})$ is identified. Thus the joint distribution of

$$(\gamma_1,\gamma_2) = \left(\frac{\pi_{13}}{\pi_{12}}\frac{\beta_1}{\beta_2}, \frac{\pi_{22}}{\pi_{12}}\frac{\beta_1}{\beta_2}\right)$$

is identified.

It remains to be shown that such (s_1,s_2,t_1,t_2) exist. Let $s_1 = x_2/t_2$ and $s_2 = x_3/t_1$, for (t_1,t_2) nonzero, to be defined shortly. This choice of (s_1,s_2) ensures that $x_2 = s_1t_2$ and $x_3 = s_2t_1$. We now must pick $t_1,t_2\in\mathbb{R}$ to satisfy

$$\begin{aligned}
x_1 &= s_1t_1+s_2t_2\frac{\beta_2}{\beta_1} \\
&= x_2\frac{t_1}{t_2}+\frac{\beta_2}{\beta_1}x_3\frac{t_2}{t_1}.
\end{aligned}$$

Equivalently, our choice of t_1,t_2 must satisfy

$$0 = (-x_1)t_1t_2+(x_2)t_1^2+\left(\frac{\beta_2}{\beta_1}x_3\right)t_2^2.$$

For any fixed t_2, this is a quadratic equation in t_1, and hence its solutions are

$$t_1 = \frac{x_1 t_2}{2x_2} \pm \frac{\sqrt{t_2^2(x_1^2 - 4x_2x_3\beta_2/\beta_1)}}{2x_2}.$$

Regardless of the value of t_2, the solutions for t_1 are real if and only if $x_1^2 \geq 4x_2x_3\beta_2/\beta_1$. Since our choice of t_2 doesn't affect the existence of a real solution to t_2, it can be chosen arbitrarily; say $t_2 = 1$. The set of (x_1, x_2, x_3) for which $x_1^2 \geq 4x_2x_3\beta_2/\beta_1$ holds has positive measure. For example, if $\beta_2/\beta_1 > 0$ it includes the quadrant $\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_2 < 0, x_3 > 0\}$.

4. Next I show that (β_1, β_2) are point identified. By assumption (iv), the mean of the reduced form coefficients exists:

$$\mathbb{E}(\pi_{12}) = \mathbb{E}\left(\frac{\beta_1}{1 - \gamma_1\gamma_2}\right) = \beta_1 \mathbb{E}\left(\frac{1}{1 - \gamma_1\gamma_2}\right).$$

The term $\mathbb{E}[1/(1-\gamma_1\gamma_2)]$ is identified since the joint distribution of (γ_1, γ_2) is identified. Thus

$$\beta_1 = \frac{\mathbb{E}(\pi_{12})}{\mathbb{E}[1/(1 - \gamma_1\gamma_2)]}$$

and hence is identified. This plus identification of the ratio β_1/β_2 implies that β_2 is identified. Note that if the nonzero mean part of assumption (iv) is dropped, but we assume additionally that $\mathbb{E}(\pi_{12}^2) < \infty$, then the magnitudes $|\beta_1|$ and $|\beta_2|$ can still be identified by

$$\mathbb{E}(\pi_{12}^2) = \beta_1^2 \mathbb{E}\left(\frac{1}{(1 - \gamma_1\gamma_2)^2}\right),$$

where now we know that the expectation on the right hand side must be nonzero.

□

Proof of proposition 4. Identification of the joint distribution of $(\gamma_1\beta_2, \beta_2)$ follows from the proof of theorem 3. The result then follows by applying lemma 6. □

Proposition 5. Suppose one of the following holds.

1. $\mathbb{P}[\text{sign}(\gamma_1) \neq \text{sign}(\gamma_2) \mid X] = 1$.
2. $\mathbb{P}(|\gamma_i| < \tau_i \mid X) = 1$ for some $0 < \tau_i < 1$, for $i = 1, 2$.

Then A6.1 and A1 hold. Assumption (ii) in proposition 3 also holds.

Proof of proposition 5. Suppress conditioning on X. In all cases I will show that there is a $\tau \in (0, 1)$ such that $\mathbb{P}[\gamma_1\gamma_2 \in (1-\tau, 1+\tau)] = 0$, which is equivalent to A6.1. Moreover, note that A6.1 implies A1.

1. Since the sign of γ_1 and γ_2 are not equal with probability one, $\mathbb{P}(\gamma_1\gamma_2 < 0) = 1$. Let τ be any number in $(0, 1)$. Then $1-\tau > 0$ and so $\mathbb{P}(\gamma_1\gamma_2 \leq 1-\tau) = 1$. Hence $\mathbb{P}[\gamma_1\gamma_2 \in (1-\tau, 1+\tau)] \leq \mathbb{P}[\gamma_1\gamma_2 > 1 - \tau] = 0$. Thus A6.1 holds. Assumption (ii) holds since $\mathbb{P}(\gamma_1\gamma_2 \leq 1) = 1 \neq 1/2$. Assumption (iv) holds since $\mathbb{P}(\gamma_1\gamma_2 < 0) = 1$ implies $\mathbb{P}(1 - \gamma_1\gamma_2 > 0) = 1$ and hence $1/(1 - \gamma_1\gamma_2) > 0$ with probability one, so its mean cannot be zero. Finally, $1 - \gamma_1\gamma_2 \geq 1$ wp1 so $1/(1 - \gamma_1\gamma_2) \leq 1$ wp1. So the mean exists.

2. By assumption there are $\tau_1, \tau_2 \in (0,1)$ such that $\mathbb{P}(|\gamma_1| \leq \tau_1) = 1$ and $\mathbb{P}(|\gamma_2| \leq \tau_2) = 1$. Let $\tilde{\tau} = \max\{\tau_1, \tau_2\} < 1$. Thus the support of (γ_1, γ_2) lies within the rectangle $[-\tilde{\tau}, \tilde{\tau}]^2$, as shown in figure 6.

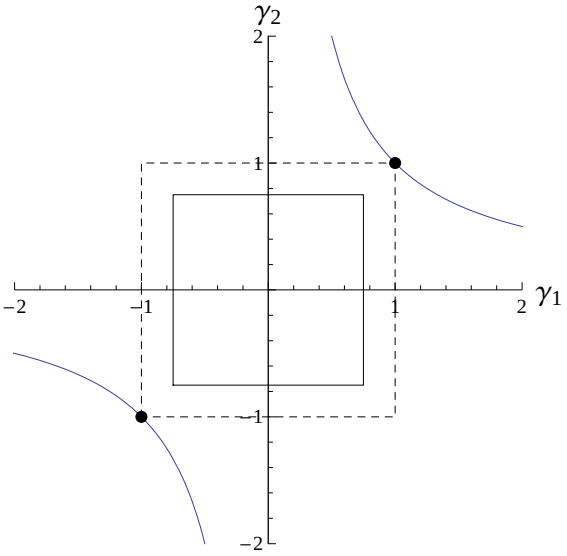


Figure 6: The solid rectangle is the boundary of $[-\tilde{\tau}, \tilde{\tau}]^2$. The dotted rectangle is the boundary of $[-1,1]^2$. The line $\gamma_1\gamma_2 = 1$ is plotted.

So $\mathbb{P}(\gamma_1\gamma_2 \leq \tilde{\tau}^2) = 1$. Let $\tau = 1 - \tilde{\tau}^2 \in (0,1)$. Then

$$\mathbb{P}(\gamma_1\gamma_2 \leq 1 - \tau) = \mathbb{P}(\gamma_1\gamma_2 \leq \tilde{\tau}^2) = 1.$$

Hence $\mathbb{P}[\gamma_1\gamma_2 \in (1-\tau, 1+\tau)] \leq \mathbb{P}[\gamma_1\gamma_2 > 1-\tau] = 0$. Thus A6.1 holds. Assumption (ii) holds since $\mathbb{P}(\gamma_1\gamma_2 \leq 1) \geq \mathbb{P}(\gamma_1\gamma_2 \leq 1-\tau) = 1 \neq 1/2$. Assumption (iv) holds since $\mathbb{P}(\gamma_1\gamma_2 \leq \tilde{\tau}^2) = 1$ and $\tilde{\tau}^2 < 1$ implies $\mathbb{P}(1 - \gamma_1\gamma_2 > 0) = 1$ and hence $1/(1-\gamma_1\gamma_2) > 0$ with probability one, so its mean cannot be zero. Finally, $1 - \gamma_1\gamma_2 \geq \tau$ wp1 implies $1/(1-\gamma_1\gamma_2) \leq 1/\tau$ so the mean exists.

□

Proof of theorem 5. I first outline the main argument, and then provide the formal justification for each step at the end. The system

$$Y_i = \frac{\gamma_i}{N-1} \sum_{j \neq i} Y_j + \beta_i Z_i + U_i,$$

for $i = 1, \ldots, N$, can be written in matrix form as

$$Y = \Gamma Y + BZ + U,$$

where

$$\Gamma = \begin{pmatrix} 0 & \frac{\gamma_1}{N-1} & \cdots & \cdots & \frac{\gamma_1}{N-1} \\ \frac{\gamma_2}{N-1} & 0 & \frac{\gamma_2}{N-1} & \cdots & \frac{\gamma_2}{N-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \frac{\gamma_{N-1}}{N-1} \\ \frac{\gamma_N}{N-1} & \frac{\gamma_N}{N-1} & \cdots & \frac{\gamma_N}{N-1} & 0 \end{pmatrix}$$

and

$$B = \begin{pmatrix} \beta_1 & 0 & \cdots & 0 \\ 0 & \beta_2 & \vdots & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \beta_N \end{pmatrix}.$$

The reduced form system is

$$Y = \widetilde{\Gamma}^{-1} B Z + \widetilde{\Gamma}^{-1} U,$$

where

$$\widetilde{\Gamma} \equiv I - \Gamma = \begin{pmatrix} 1 & -\frac{\gamma_1}{N-1} & \cdots & \cdots & -\frac{\gamma_1}{N-1} \\ -\frac{\gamma_2}{N-1} & 1 & -\frac{\gamma_2}{N-1} & \cdots & -\frac{\gamma_2}{N-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & -\frac{\gamma_{N-1}}{N-1} \\ -\frac{\gamma_N}{N-1} & -\frac{\gamma_N}{N-1} & \cdots & -\frac{\gamma_N}{N-1} & 1 \end{pmatrix}.$$

The inverse of this matrix can be written as

$$\widetilde{\Gamma}^{-1} = \frac{C'}{\det(\widetilde{\Gamma})}$$

where C is the matrix of cofactors. The key observation for the proof is that the rows of $\widetilde{\Gamma}$ each depend on a single random variable, γ_i for row i. Consider the vector of coefficients on Z_i. This is the ith column of $\widetilde{\Gamma}^{-1}B$. The element on the kth row of the ith column is

$$(\widetilde{\Gamma}^{-1})_{ki} = \frac{1}{\det(\widetilde{\Gamma})}(C')_{ki} = \frac{1}{\det(\widetilde{\Gamma})}(C)_{ik} = \frac{1}{\det(\widetilde{\Gamma})}(-1)^{i+k} M_{ik},$$

where M_{ik} is the (i,k)th-minor, the determinant of the matrix obtained by deleting row i and column k of $\widetilde{\Gamma}$. Let

$$\Pi = \widetilde{\Gamma}^{-1} B.$$

Then

$$\pi_{ki} \equiv (\Pi)_{ki} = \det(\widetilde{\Gamma})^{-1} (-1)^{i+k} M_{ik} \beta_i$$

is the coefficient on Z_i in the kth equation. By the same argument as in the two equation case, the joint distribution of reduced form coefficients $(\pi_{1i}, \ldots, \pi_{Ni})$ on the ith instrument is point identified, for all $i = 1, \ldots, N$.

By the structure of $\widetilde{\Gamma}$, when row i is deleted, γ_i no longer appears in the remaining submatrix.

Consequently, except for inside the determinant term, every reduced form coefficient on Z_i depends only on the $N-1$ random coefficients $\{\gamma_k : k \neq i\}$.

Thus by dividing the coefficient on Z_i in the ith equation, π_{ii}, into the coefficient on Z_i in all other equations $k \neq i$, π_{ki}, the determinant and β_i terms cancel, since they are common to all coefficients, and we obtain an $(N-1)$-dimensional random vector which is a function of the $N-1$ structural random coefficients $\{\gamma_k : k \neq i\}$:

$$\left(\frac{\pi_{1i}}{\pi_{ii}}, \ldots, \frac{\pi_{ki}}{\pi_{ii}}, \ldots, \frac{\pi_{Ni}}{\pi_{ii}}\right),$$

where $k = i$ is not included, and

$$\frac{\pi_{ki}}{\pi_{ii}} = \frac{(-1)^{i+k} M_{ik}}{(-1)^{2i} M_{ii}} \quad \text{for } k = 1, \ldots, N,\ k \neq i. \tag{10}$$

Temporarily thinking of the reduced form parameters as constants, equation (10) is a system of $(N-1)$ equations in $(N-1)$ unknowns, $\{\gamma_k : k \neq i\}$. The unique solution to this system of equations is

$$\gamma_k = \frac{(N-1)(\pi_{ki}/\pi_{ii})}{1 + \sum_{j \neq k, j \neq i} (\pi_{ji}/\pi_{ii})}$$

for $k = 1, \ldots, N$, $k \neq i$. This mapping from the parameters $\{\pi_{ki}/\pi_{ii} : k \neq i\}$ to $\{\gamma_k : k \neq i\}$ is one-to-one and differentiable and hence the joint distribution of $\{\gamma_k : k \neq i\}$ can be written in terms of the joint distribution of $\{\pi_{ki}/\pi_{ii} : k \neq i\}$ via the change of variables formula (e.g., Munkres (1991) theorem 17.2). Hence the joint distribution of $\{\gamma_k : k \neq i\}$ is point identified.

The same argument can be applied to the coefficients on Z_j for any $j \neq i$ to obtain the joint distribution of $\{\gamma_i : i \neq j\}$, which concludes the main outline of the proof.

The proof is finished by providing formal justification for the steps above. I will show that

1. The reduced form matrix is invertible with probability 1.

2. The distribution of reduced form parameters satisfy the moment conditions needed to apply lemma 2 to identify the distribution of reduced form parameters in the single equation model for the linear combination $t_1 Y_1 + \cdots + t_N Y_N$, where $t_1, \ldots, t_N \in \mathbb{R}$.

3. The diagonal elements π_{ii} are nonzero with probability one, so that the ratio random variables π_{ki}/π_{ii} are well-defined.

4. The denominator of the mapping from the ratios of the reduced form coefficients to the structural parameters is bounded away from zero with probability one, which both ensures that this unique solution to the system (10) exists and that the mapping is differentiable on its domain, which is sufficient to apply the change-of-variables theorem, since the mapping is rational (the ratio of two polynomials) and hence is differentiable everywhere where the denominator is not zero.

Let $\|\cdot\|_\infty$ be the maximum row-sum matrix norm:

$$\|A\|_\infty \equiv \max_{1 \leq i \leq L} \sum_{j=1}^{L} |a_{ij}|.$$

For the ith row of Γ,

$$\begin{aligned}\sum_{j=1}^{L} |(\Gamma)_{ij}| &\leq \left(\frac{\tau}{N-1} + \cdots + \frac{\tau}{N-1}\right) \\ &= \tau\end{aligned}$$

where the first line follows since $|\gamma_i| \leq \tau$ for all i, and the last line follows since we're summing up $N-1$ different terms. Hence $\|\Gamma\|_\infty \leq \tau < 1$. Thus lemma 9 implies that $I - \Gamma$ is invertible. Hence $\mathbb{P}(\det(I-\Gamma) = 0) = 0$. Next,

$$\begin{aligned}\|(I-\Gamma)^{-1}\|_\infty &\leq \frac{1}{1 - \|\Gamma\|_\infty} \\ &\leq \frac{1}{1-\tau} \\ &< \infty.\end{aligned}$$

The first line follows by the third exercise following corollary 5.6.16 on page 351 of Horn and Johnson (2013). The second follows since $\|\Gamma\|_\infty \leq \tau$. Since we're using the maximum row-sum norm, this implies that the absolute value of each element of $(I-\Gamma)^{-1}$ is bounded. Hence the reduced form coefficients are bounded and hence all of their moments exist and their distribution is uniquely determined by these moments.

Next we consider the structure of the matrix of reduced form coefficients, $(I - \Gamma)$. It is helpful to derive the results for the slightly more general matrix

$$A_n = \begin{pmatrix} 1 & -a_1 & \cdots & -a_1 \\ -a_2 & 1 & \ldots & -a_2 \\ \vdots & \vdots & \ddots & \vdots \\ -a_n & -a_n & \cdots & 1 \end{pmatrix}$$

with the main case of interest being $a_k = \gamma_k/(N-1)$ and $n = N$. As a running example, consider

$$A_3 = \begin{pmatrix} 1 & -a_1 & -a_1 \\ -a_2 & 1 & -a_2 \\ -a_3 & -a_3 & 1 \end{pmatrix}.$$

The determinant of A_n is

$$\det(A_n) = 1 - \left(\sum_{i_1<i_2} a_{i_1}a_{i_2} + 2\sum_{i_1<i_2<i_3} a_{i_1}a_{i_2}a_{i_3} + \cdots + (n-1)\sum_{i_1<\cdots<i_n} a_{i_1}\cdots a_{i_n}\right).$$

The first sum on the right hand side is the sum of all possible products of two elements from $\{a_1, \ldots, a_n\}$ (where order does not matter). The second sum is the the sum of all possible products of three elements from $\{a_1, \ldots, a_n\}$. Likewise for the rest of the sums. For example,

$$\det(A_3) = 1 - a_1a_2 - a_1a_3 - a_2a_3 - 2a_1a_2a_3.$$

The diagonal element $[A_n^{-1}]_{i,i}$ is the determinant of the submatrix of A_n with row i and column i deleted, divided by the determinant of A_n. This submatrix has the same form as A_n and hence

its determinant has the same form, just with the element a_i omitted from all summations. For example,

$$\det(A_3)[A_3^{-1}]_{1,1} = 1 - a_2 a_3.$$

The off diagonal elements $[A_n^{-1}]_{i,j}$, $i \neq j$, have a similar structure:

$$\det(A_n)[A_n^{-1}]_{i,j} = a_i \left(1 + \sum_{k:k\neq i,j} a_k + \sum_{k_1<k_2:k_1,k_2\neq i,j} a_{k_1} a_{k_2} + \cdots + \sum_{k_1<\cdots<k_{n-2}:k_1,\ldots,k_{n-2}\neq i,j} a_{k_1} \cdots a_{k_{n-2}} \right).$$

The first sum on the right hand side is the sum of all elements from $\{a_1, \ldots, a_n\} \setminus \{a_i, a_j\}$. The second sum is the sum of all products of two elements from $\{a_1, \ldots, a_n\} \setminus \{a_i, a_j\}$ (where the order of the product doesn't matter). Likewise up through the last sum. For example,

$$\det(A_3)[A_3^{-1}]_{2,1} = a_2(1 + a_3).$$

Using these formulas for the elements of A_n^{-1}, for any i and k, $k \neq i$,

$$\begin{aligned}
&\det(A_n) \left([A_n^{-1}]_{ii} + \sum_{j\neq k, j\neq i} [A_n^{-1}]_{ji} \right) \\
&= 1 + \sum_{\ell:\ell\neq i,k} a_\ell + \sum_{\ell_1<\ell_2:\ell_1,\ell_2\neq i,k} a_{\ell_1} a_{\ell_2} + \cdots + \sum_{\ell_1<\cdots<\ell_{n-2}:\ell_1,\ldots,\ell_{n-2}\neq i,k} a_{\ell_1} \cdots a_{\ell_{n-2}} \\
&= \frac{1}{a_k} \det(A_n)[A_n^{-1}]_{ki}.
\end{aligned}$$

For example, for $n = 3$, $i = 1$ and $k = 2$,

$$\begin{aligned}
\det(A_3) \left([A_3^{-1}]_{11} + \sum_{j\neq 2, j\neq 1} [A_n^{-1}]_{ji} \right) &= \det(A_3)([A_3^{-1}]_{11} + [A_3^{-1}]_{31}) \\
&= \det(A_3)([1 - a_2 a_3] + [a_3(1 + a_2)]) \\
&= \det(A_3)(1 + a_3) \\
&= \frac{1}{a_2} \det(A_3) a_2 (1 + a_3) \\
&= \frac{1}{a_2} \det(A_3)[A_3^{-1}]_{21}.
\end{aligned}$$

Hence for any i and k, $k \neq i$,

$$a_k = \frac{\det(A_n)[A_n^{-1}]_{ki}}{\det(A_n) \left([A_n^{-1}]_{ii} + \sum_{j\neq k, j\neq i} [A_n^{-1}]_{ji} \right)}$$

which is the same form of the mapping from the reduced form coefficients to the structural coefficients given above, where note that we can divide both the numerator and denominator by $[A_n^{-1}]_{ii}$ to put the right hand side in terms of ratios of off-diagonal elements to diagonal elements, and we used the notation

$$\pi_{ij} = [A_n^{-1}]_{ij}$$

and

$$a_k = \frac{\gamma_k}{N-1}.$$

This shows how to derive the mapping from the elements of the inverse matrix to the elements of the original matrix (i.e., from the reduced form parameters to the structural coefficients). Next, letting $n = N$ and noting that $a_k = \gamma_k/(n-1) \leq \tau/(n-1)$ the numerator of the ith diagonal component of A_n^{-1} is

$$\begin{aligned}
\det(A_n)\pi_{ii} &= \det(A_n)[A_{n-1}^{-1}]_{ii} \\
&\geq 1 - \left(\binom{n-1}{2} \left(\frac{\tau}{n-1}\right)^2 + 2\binom{n-1}{3} \left(\frac{\tau}{n-1}\right)^3 + \cdots + ([n-1]-1)\binom{n-1}{n} \left(\frac{\tau}{n-1}\right)^{n-1} \right) \\
&= 1 - \sum_{k=2}^{n-1} \binom{n-1}{k} (k-1) \left(\frac{\tau}{n-1}\right)^k \\
&= \binom{n-1}{0} \left(\frac{\tau}{n-1}\right)^0 + \sum_{k=2}^{n-1} \binom{n-1}{k} \left(\frac{\tau}{n-1}\right)^k - \sum_{k=2}^{n-1} k\binom{n-1}{k} \left(\frac{\tau}{n-1}\right)^k \\
&= \sum_{k=0}^{n-1} \binom{n-1}{k} \left(\frac{\tau}{n-1}\right)^k - 1\binom{n-1}{1} \left(\frac{\tau}{n-1}\right)^1 - \sum_{k=2}^{n-1} k\binom{n-1}{k} \left(\frac{\tau}{n-1}\right)^k \\
&= \sum_{k=0}^{n-1} \binom{n-1}{k} \left(\frac{\tau}{n-1}\right)^k - \sum_{k=0}^{n-1} k\binom{n-1}{k} \left(\frac{\tau}{n-1}\right)^k \\
&> 0
\end{aligned}$$

where in the first line A_{n-1} is stands for the submatrix of A_n with the ith row and column of A_n deleted, and the last line follows since

$$\sum_{k=0}^{n} k\binom{n}{k} \left(\frac{\tau}{n}\right)^k < \sum_{k=0}^{n} \binom{n}{k} \left(\frac{\tau}{n}\right)^k$$

holds for all $\tau \in (0,1)$ and all n (it's important here that the indexing starts at $k = 0$). This argument also shows that $\det(A_n) > 0$. These statements all hold with probability one over the distribution of the γ_i's. Thus $\mathbb{P}(\pi_{ii} > 0) = 1$ (and in fact we've shown that it's actually strictly bounded away from zero).

Finally, consider the denominator

$$\begin{aligned}
&\det(A_n) \left([A_n^{-1}]_{ii} + \sum_{j \neq k, j \neq i} [A_n^{-1}]_{ji} \right) \\
&= 1 + \sum_{\ell : \ell \neq i,k} a_\ell + \sum_{\ell_1 < \ell_2 : \ell_1, \ell_2 \neq i,k} a_{\ell_1} a_{\ell_2} + \cdots + \sum_{\ell_1 < \cdots < \ell_{n-2} : \ell_1, \ldots, \ell_{n-2} \neq i,k} a_{\ell_1} \cdots a_{\ell_{n-2}}.
\end{aligned}$$

The domain of this function is $[-\tau/(n-1), \tau/(n-1)]^{n-2}$, $\tau \in (0,1)$. This function is strictly increasing in each component over this domain. Hence it is minimized at $a_\ell = -\tau/(n-1)$, which

gives

$$\begin{aligned}
\det(A_n)\left([A_n^{-1}]_{ii} + \sum_{j\neq k, j\neq i}[A_n^{-1}]_{ji}\right) &\geq 1 + \sum_{k=1}^{n-2}\binom{n-2}{k}\left(\frac{\tau}{n-1}\right)^k(-1)^k \\
&= 1 + \sum_{k=1}^{n-2}\binom{n-2}{k}\left(\frac{-\tau}{n-1}\right)^k \\
&= 1 + \sum_{k=0}^{n-2}\binom{n-2}{k}\left(\frac{-\tau}{n-1}\right)^k - 1 \\
&= \left(1 - \frac{\tau}{n-1}\right)^{n-2} \\
&> 0
\end{aligned}$$

where the fourth line follows by the binomial theorem and the last line since $\tau/(n-1) < 1$. Hence the denominator of our mapping from reduced form coefficients to structural coefficients is strictly positive and bounded away from zero with probability 1. □

Lemma 9. Let Γ be such that $\|\Gamma\| < 1$, where $\|\cdot\|$ is any matrix norm. Then $(I - \Gamma)$ is invertible, $\sum_{k=0}^{\infty}\Gamma^k$ converges, and

$$(I-\Gamma)^{-1} = \sum_{k=0}^{\infty}\Gamma^k.$$

Proof of lemma 9. See the first exercise following corollary 5.6.16 on page 351 of Horn and Johnson (2013). □

Proof of theorem 6. This result follows from a minor generalization of the proof of theorem 5. Recall that $N_j = |\mathscr{N}(j)|$ and $1_{ji} = \mathbb{1}[i \in \mathscr{N}(j)]$. Our structural system can be written as

$$Y = \Gamma Y + BZ + U,$$

where

$$\Gamma = \begin{pmatrix}
0 & \frac{\gamma_1}{N_1}1_{12} & \cdots & \cdots & \frac{\gamma_1}{N_1}1_{1N} \\
\frac{\gamma_2}{N_2}1_{21} & 0 & \frac{\gamma_2}{N_2}1_{23} & \cdots & \frac{\gamma_2}{N_2}1_{2N} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\vdots & \vdots & \cdots & \ddots & \frac{\gamma_{N-1}}{N_{N-1}}1_{N-1,N} \\
\frac{\gamma_N}{N_N}1_{N,1} & \frac{\gamma_N}{N_N}1_{N,2} & \cdots & \frac{\gamma_N}{N_N}1_{N,N-1} & 0
\end{pmatrix}.$$

This is basically the vector

$$\begin{pmatrix} \gamma_1/N_1 \\ \vdots \\ \gamma_N/N_N \end{pmatrix}$$

replicated N times and then element-wise multiplied against the adjacency matrix. In general, for $j = 1, \ldots, N$, $j \neq i$,

$$\gamma_j = \frac{N_j(\pi_{ji}/\pi_{ii})}{1_{ji} + \sum_{k\neq j, k\neq i}1_{jk}(\pi_{ki}/\pi_{ii})}.$$

In the classical linear-in-means case, this formula simplifies to the one obtained in the proof of theorem 5 since $1_{ji} = 1$ for all distinct i, j. The assumption that $N_j \geq 1$ with probability one ensures that this denominator is never zero. This assumption also ensures that for person j, there is some other person i who directly affects j. This implies that π_{ji}, the reduced form effect of person i's covariate Z_i on person j, is nondegenerate, which is necessary to recover γ_i. Note that in the linear-in-means case, π_{ji} nondegenerate is guaranteed by the assumption that everyone in the reference group affects j. The remainder of the proof follows as in the proof of theorem 5. $\square$