ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Masten, Matthew

Working Paper

Random coefficients on endogenous variables in simultaneous equations models

cemmap working paper, No. CWP01/14

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Masten, Matthew (2014) : Random coefficients on endogenous variables in simultaneous equations models, cemmap working paper, No. CWP01/14, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2013.0114

This Version is available at:
https://hdl.handle.net/10419/97411

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

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 CWP01/14

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

December 29, 2013

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. 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 requires full support instruments, but allows for nearly arbitrary distributions of unobservables. The second allows for continuous instruments without full support, but places tail restrictions on the distributions of unobservables. I show that a nonparametric sieve maximum likelihood estimator for these distributions is consistent. I apply my results to the Add Health data to analyze the social determinants of obesity.

*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, Rosa Matzkin, Konrad Menzel, and Alex Torgovitsky 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, and the University of Iowa. 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.

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. 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 $E(\gamma_1 \mid X)$ and $E(\gamma_2 \mid X)$, 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. Even so, I prove that the 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 that, with some restrictions on the distribution of unobservables, full support Z can be relaxed. I propose a consistent nonparametric estimator for the distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$.

Throughout I assume all coefficients on exogenous variables are also random. Note that the ad-

[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).

ditive 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.

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) and obesity (Christakis and Fowler 2007, Cohen-Cole and Fletcher 2008). In an empirical application, I revisit the controversial topic of the social determinants of obesity. I use the Add Health data to construct best friend pairs. I set the outcomes Y_1 and Y_2 to be the change in each friends' weight between two survey waves, and I choose Z_i to be the change in person i's height over the same time period. 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 equal to or larger 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. 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 13).

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 $E(\gamma_1)$ 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), Bajari, Fox, Kim, and Ryan (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 2012). 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 2010, and Berry and Haile 2011), 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. Kasy (2013) 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 discuss later, the entire joint distribution of unobservables in (1) is unlikely to be 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

$$Y_1 = \frac{1}{1-\gamma_1\gamma_2}\left[U_1 + \gamma_1 U_2 + \beta_1 Z_1 + \gamma_1\beta_2 Z_2 + \delta_1' X + \gamma_1\delta_2' X\right] \tag{2}$$
$$Y_2 = \frac{1}{1-\gamma_1\gamma_2}\left[\gamma_2 U_1 + U_2 + \gamma_2\beta_1 Z_1 + \beta_2 Z_2 + \gamma_2\delta_1' X + \delta_2' X\right].$$

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.

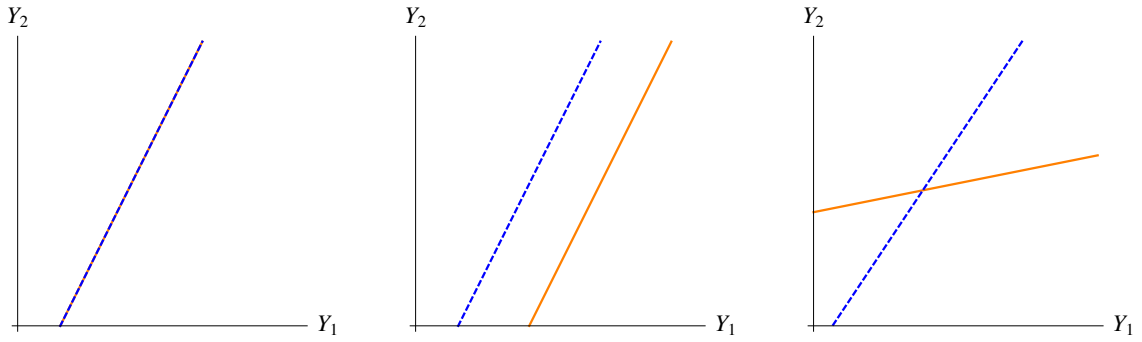


Figure 1: These figures plot the lines $Y_1 = \gamma_1 Y_2 + C_1$, shown as the solid line, and $Y_2 = \gamma_2 Y_1 + C_2$, shown as the dashed line. By varying γ_1, γ_2, C_1, and C_2, each plot shows a different possible configuration of the system: parallel and overlapping, parallel and nonoverlapping, 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 incomplete without further assumptions (see Tamer 2003 and Lewbel 2007 for a discussion of complete and incomplete models). To ensure 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). $P(\gamma_1\gamma_2 = 1 \mid X, Z) = 0$.

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 (2). 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 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 $P(\gamma_1\gamma_2 = 1 \mid X, Z) < p$ 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 (2). 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 $E(Y_1 \mid X, Z)$ and $E(Y_2 \mid X, Z)$, without further restrictions on the distribution of (γ_1, γ_2).

In the next section, 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

[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.

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. □

3 Identification

In this section I prove two point-identification results for system (1), neither of which require parametric assumptions on the distribution of coefficients. In section 3.1, I consider identification with no assumptions on the distribution of unobservables $(U_1, U_2, \gamma_1, \gamma_2)$ beyond the unique solution assumption A1. Consequently, I impose strong assumptions on the instruments to achieve full point identification: I require Z to be independent of all unobservables and have full support, given X. In section 3.2, I relax the full support assumption but impose additional restrictions on the distribution of $(U_1, U_2, \gamma_1, \gamma_2)$. Specifically, I require Z to have continuous variation and I require the reduced form coefficients to have finite moments which uniquely determine their distribution. In both sections I show that the marginal distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$ are point identified.

Throughout the paper I use the following notation and definition of identification. Let $F_{B,D,U,\Gamma|X}$ denote the joint distribution of $(\beta_1, \beta_2, \delta_1, \delta_2, U_1, U_2, \gamma_1, \gamma_2)$ given X. Let $\alpha = F_{B,D,U,\Gamma|X}$ denote this unknown structural distribution. Let $\mathscr{A}$ denote the set of α's which satisfy the assumptions of the model under consideration. Finally, let $F(\cdot \mid \alpha)$ denote the distribution of observables (Y, X, Z) when the parameter is α, and let $F_{\text{OBS}}(\cdot)$ denote the observed distribution of (Y, X, Z) in the data.

Definition (Identification)**.** The parameters α and $\tilde{\alpha}$ are *observationally equivalent* if $F(\cdot \mid \alpha) = F(\cdot \mid \tilde{\alpha})$. The *identification region* for α is the set $\mathscr{A}_I \equiv \{\alpha \in \mathscr{A} : F(\cdot \mid \alpha) = F_{\text{OBS}}(\cdot)\}$. If this region is a singleton, α is *point identified*. If this region is a strict subset of $\mathscr{A}$, but is not a singleton, α is *partially identified*. Otherwise, α is *not identified*. A *feature* of α is a known function $C : \mathscr{A} \to \mathscr{C}$, where $\mathscr{C}$ is some space of interest. For example, $C(\alpha)$ may denote the distribution of $\gamma_1 \mid X$. Two values c and $\tilde{c}$ of a feature C are observationally equivalent if there exist two observationally equivalent α and $\tilde{\alpha}$ with $c = C(\alpha)$ and $\tilde{c} = C(\tilde{\alpha})$. The identification region for the values of the feature C is the set $\{C(\alpha) \in \mathscr{C} : \alpha \in \mathscr{A}_I\}$. If this region is a singleton, the value of the feature C is point identified. If this region is a strict subset of $\mathscr{C}$, but is not a singleton, the value of the feature C is partially identified. Otherwise, the value of the feature C is not identified.

For a fixed $X = x$, $F_{B,D,U,\Gamma|X=x}$ is a $6 + 2K$ dimensional distribution. Even with no covariates X, so that $K = 0$, this is a 6 dimensional distribution, while the data (Y_1, Y_2, Z_1, Z_2) has only 4 dimensions. Consequently, it is unlikely that we can obtain point identification of an arbitrary joint

distribution $F_{B,D,U,\Gamma|X}$ without restricting its dimension. I therefore focus on providing sufficient conditions under which we can still obtain point identification of particular features of α. These conditions continue to allow for high dimensional unobservables.

Throughout this paper, 'identified' means 'point identified'. 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 Instruments with full support

In this section I provide conditions under which the marginal distributions $\gamma_1 \mid X$ and $\gamma_2 \mid X$ are identified, even if the reduced form mean regression fails to exist because the structural equations are nearly parallel too often.

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

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 $P(\beta_1 = 0 \mid X) > 0$, then we can identify the distribution of γ_2 conditional on X and $\beta_1 \neq 0$. Likewise, if $P(\beta_2 = 0 \mid X) > 0$, 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, $P(\beta_1 = 0 \mid X) > 0$ 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 $P(\beta_2 = 0 \mid X) > 0$ 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] 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.

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

[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.

This assumption is key in proving identification with minimal restrictions on the distribution of the unobservables. In section 3.2, I relax this assumption, at the price of placing additional restrictions on the distribution 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′ in the next section), then I show one can identify the distribution of social interaction effects with experimental data.*

Theorem 1. Under A1, A2, A3, and A4, the conditional distributions $\gamma_1 \mid X = x$ and $\gamma_2 \mid X = x$ are identified for each $x \in \text{supp}(X)$.

While I gather all proofs in appendix A, I sketch the proof of theorem 1 here to show its main idea. Fix a value of $x \in \text{supp}(X)$. The reduced form system (2) is

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

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 using a result on identification of random coefficients in single equation models (see lemma 1 below), we can identify the joint distribution of

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

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

$$(\pi_{13}, \pi_{23}) = \left(\frac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2}, \frac{\beta_2}{1 - \gamma_1 \gamma_2} \right) \tag{3}$$

and from this we have $\gamma_1 = \pi_{13}/\pi_{23}$. Similarly for 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 following lemma about single-equation random coefficient models is a key step in the proof of theorem 1.

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.

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) 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:

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

where the second line follows since $Z \perp\!\!\!\perp (A, B)$ and by the definition of the characteristic function for (A, B). 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).

Beyond the unique solution assumption A1, no restrictions on the distribution of (B, D, U, Γ) are required for identification of the distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$. Specifically, the unobservable variables can be arbitrarily dependent.

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.*

Theorem 1 provides a result for the marginal distributions of endogenous variable random coefficients. As mentioned earlier, it is unlikely that we will be able to obtain full point identification of the distribution of all unobservables. Specifically, identification of the joint distribution of all structural parameters obtains if the joint distribution of *all* reduced form coefficients (π_1, π_2) obtains. This latter identification result, however, intuitively requires independent variation of all the regressors in the reduced form system, which is not possible since all instruments enter each reduced form equation and thus changing a single variable necessarily affects both equations. See the remark on page 13 for further discussion. Even so, it may be possible to obtain point identification of other functionals of this distribution, such as the joint distribution of γ_1 and γ_2. Such a result would, for example, allow us to learn whether assortative matching between friends occurred along the dimension of social susceptibility. If one of β_1 or β_2 were constant, then identification of

this joint distribution would obtain from equation (3) via a change of variables. This argument, however, does not work when β_1 and β_2 are random. I leave the general question of what additional functionals of the full joint distribution of unobservables are identified to future work; although note that in some cases the setting naturally provides additional restrictions on this joint distribution, 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.*

The following result uses the proof of theorem 1 to examine triangular systems, a case of particular relevance for the literature on heterogeneous treatment effects.

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

$$\begin{aligned} Y_1 &= \gamma_1 Y_2 + \delta_1' X + U_1 \\ Y_2 &= \beta_2 Z_2 + \delta_2' X + U_2. \end{aligned} \tag{4}$$

Suppose the assumptions of either theorem 1 or 2 hold. Then the joint distribution of $(\gamma_1, \beta_2) \mid X$ is identified.

For example, suppose Y_1 is log-wage and Y_2 is education. While the 2SLS estimator of γ_1 in the triangular model (4) 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 1 makes no restrictions on the dependence structure of the unobservables $(U_1, U_2, \gamma_1, \beta_2, \delta_1, \delta_2)$, which allows (4) 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. Hoderlein et al. (2010, page 818) also discuss identification of a triangular model like (4), but they assume β_2 is constant.

Remark 2 (The role of additive separability and linearity)**.** In both systems (1) and (4), 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. □

I conclude this section with several remarks on the related literature and by noting that the identification strategy does not easily generalize to systems with more than two equations. Essentially, the inverse of the matrix of random coefficients on Y becomes too unwieldy. For example, consider the three-equation system

$$\begin{aligned} Y_1 &= \gamma_{12}Y_2 + \gamma_{13}Y_3 + \beta_1 Z_1 + U_1 \\ Y_2 &= \gamma_{21}Y_1 + \gamma_{23}Y_3 + \beta_2 Z_2 + U_2 \\ Y_3 &= \gamma_{31}Y_1 + \gamma_{32}Y_2 + \beta_3 Z_3 + U_3. \end{aligned}$$

The reduced form equation for Y_1 is

$$\begin{aligned} Y_1 = \det(I - \Gamma)^{-1}[&(1 - \gamma_{23}\gamma_{32})\beta_1 Z_1 + (\gamma_{12} + \gamma_{13}\gamma_{32})\beta_2 Z_2 + (\gamma_{13} + \gamma_{12}\gamma_{23})\beta_3 Z_3 \\ &+ (1 - \gamma_{23}\gamma_{32})U_1 + (\gamma_{12} + \gamma_{13}\gamma_{32})U_2 + (\gamma_{13} + \gamma_{12}\gamma_{23})U_3], \end{aligned}$$

and similarly for Y_2 and Y_3. The cross equation reduced form coefficients on each instrument can be identified under assumptions like those above, but they are complicated functions of the structural distributions we wish to identify. This precludes simple cancellations as in theorem 1. Moreover, these three dimensional random vectors are functions of the six dimensional vector of endogenous variable coefficients, as well as the coefficient on the instrument. One possible approach is to assume some of the coefficients are known to be zero a priori. This would not necessarily eliminate all simultaneity, and thus may still be of interest. For example, if $\gamma_{23} = \gamma_{32} = 0$ is known a priori, then a similar argument to that above suggests that the marginal distributions of γ_{12} and γ_{13} are identified, by comparing the coefficients on the equation for Y_1. This situation can arise if this system represents a social network, in which case assuming the existence of intransitive triads implies certain zero restrictions (e.g. Bramoulle, Djebbari, and Fortin 2009). Nonetheless, I do not pursue this idea here.

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. Although that proposition applies to systems of equations, the equations in those

systems are not allowed to have common regressors, which rules out fully simultaneous equations models, as well as triangular models. To see this, consider a simple two equation system

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} + \begin{pmatrix} Z_1 & 0 \\ 0 & Z_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},$$

written in the form of Beran and Millar's equation (1.1). As above, $(U_1, U_2, \beta_1, \beta_2)$ are unobserved random variables. Beran and Millar's identification result, proposition 2.2, makes a support assumption, labeled 2.1: they require the support of the vector $(t_1 Z_1, t_2 Z_2)$ to contain an open set in $\mathbb{R}^2$ for each $(t_1, t_2) \in \mathbb{R}^2$, t_1, t_2 nonzero. This cannot hold if Z_1 and Z_2 are functionally related, such as when $Z_1 = Z_2$. Intuitively, when $Z_1 = Z_2$, we cannot independently vary the regressor in the first equation from the regressor in the second equation, which precludes learning the joint distribution of (β_1, β_2).

This common regressor issue occurs by construction in a simultaneous equations model. Writing our system of reduced form equations (2) in the form of Beran and Millar's equation (1.1) we have

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \dfrac{U_1 + \gamma_1 U_2 + \delta_1' X + \gamma_1 \delta_2' X}{1 - \gamma_1 \gamma_2} \\ \dfrac{\gamma_2 U_1 + U_2 + \gamma_2 \delta_1' X + \delta_2' X}{1 - \gamma_1 \gamma_2} \end{pmatrix} + \begin{pmatrix} Z_1 & Z_2 & 0 & 0 \\ 0 & 0 & Z_1 & Z_2 \end{pmatrix} \begin{pmatrix} \dfrac{\beta_1}{1 - \gamma_1 \gamma_2} \\ \dfrac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2} \\ \dfrac{\gamma_2 \beta_1}{1 - \gamma_1 \gamma_2} \\ \dfrac{\beta_2}{1 - \gamma_1 \gamma_2} \end{pmatrix}.$$

Conditional on X, Beran and Millar's support condition is then that the support of $(t_1 Z_1,\ t_1 Z_2,\ t_2 Z_1,\ t_2 Z_2)$ contains an open ball in $\mathbb{R}^4$ for all $(t_1, t_2) \in \mathbb{R}^2$, t_1, t_2 nonzero. This does not hold. Essentially, simultaneity implies that each instrument necessarily enters all reduced form equations. Consequently, we will not be able to independently vary the regressors across equations to learn the joint distribution of all reduced form coefficients.

For the simultaneous equations model considered here, and conditional on X, Hahn's equation (1) writes the model as

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

His support condition (assumption v) then assumes the support of $t_1 + t_2 Z_1 + t_3 Z_2$ contains an open ball in $\mathbb{R}$ for all nonzero $(t_1, t_2, t_3) \in \mathbb{R}^3$. This support assumption is not sufficient for Beran and Millar's support assumption. Moreover, as shown above, Beran and Millar's support assumption cannot hold in a simultaneous equations model by construction. 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. □

3.2 Instruments with continuous variation

In this section, I show that we can relax the full support assumption on Z to just requiring that Z has continuous variation. The trade-off is that I place restrictions on the distribution of random coefficients.

Assumption A5 (Moment determinacy).

1. Conditional on $X = x$, the absolute moments of the reduced form coefficients $\pi_i \equiv (\pi_{i1}, \pi_{i2}, \pi_{i3})$,

$$\int |p_1|^{\alpha_1} |p_2|^{\alpha_2} |p_3|^{\alpha_3} \, dF_{\pi_i|X}(p \mid x), \qquad \alpha \in \mathbb{N}^3,$$

   are finite, $i = 1, 2$, for each $x \in \text{supp}(X)$. $\mathbb{N}$ denotes the natural numbers.

2. The distribution of $\pi_i \mid X = x$ is uniquely determined by its moments, $i = 1, 2$, for each $x \in \text{supp}(X)$.

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, Bajari, Fox, Kim, and Ryan (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.

Assumption A4′ (Instruments have continuous variation)**.** $\text{supp}(Z \mid X = x)$ contains an open ball in $\mathbb{R}^2$, for each $x \in \text{supp}(X)$.

This assumption requires that there always be some region where we can vary (Z_1, Z_2) in any direction. For example, 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.

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 $x \in \text{supp}(X)$.

The proof is essentially identical to that of theorem 1. The only difference is that in the first step we apply a different identification result for the single-equation random coefficient model, described as follows.

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

For a scalar Z, this result was proved in Beran's (1995) proposition 2. Lemma 2 here shows that the result holds for any finite dimensional vector Z, as needed for the simultaneous equations analysis, and also uses a different proof technique. The proof 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. 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 $E(\gamma_1 \mid X)$ and $E(\gamma_2 \mid X)$, assuming these structural moments exist.

A sufficient condition for A5, in terms of the structural parameters, is the following.

Assumption A6 (Restrictions on structural unobservables).

1. $P(|1 - \gamma_1\gamma_2| \geq \tau \mid X) = 1$ for some $\tau > 0$. That is, $1 - \gamma_1\gamma_2$ is bounded away from zero, or equivalently, $\gamma_1\gamma_2$ is bounded away from 1.
2. $\gamma_1 \mid X$ and $\gamma_2 \mid X$ have compact support.
3. The distributions of $\beta_1 \mid X$ and $\beta_2 \mid X$ have finite absolute moments and their moment generating functions exist.
4. The distribution of $(U_1, U_2, \delta_1, \delta_2) \mid X$ has finite absolute moments. The moment generating function of $(U_1, U_2, \delta_1, \delta_2) \mid X$ exists.

Proposition 2. A6 implies A5.

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 3 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. A6.2 can be relaxed at the expense of less interpretable tail conditions; see the proof of proposition 2. A6.3 and A6.4 accommodate most well known distributions, such as the normal distribution, as well as any compactly supported distribution.

4 Estimation

In this section I consider a nonparametric sieve maximum likelihood estimator of the distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$, under the identification assumptions of either section 3.1 or section 3.2. I also discuss the two-stage least squares estimator.

A sieve approach is attractive for several reasons. Sieves allow us to easily impose additional structural assumptions. For applications to pairs of people, the labels of person 1 versus person 2 may not matter, and hence we may assume that γ_1 and γ_2 have the same distribution, and likewise for U_1 and U_2, β_1 and β_2, and δ_1 and δ_2. This assumption is easily imposed using sieves. Similarly, monotonicity assumptions like $\gamma_1 \geq 0$ can be imposed by restricting the support of the sieve estimator. Sieves can also nest parametric assumptions about the distribution of random coefficients. The sieve MLE can thus be thought of as a generalization of the classical full information maximum likelihood (FIML) estimator, allowing for random coefficients.

One complication of using sieve maximum likelihood, however, is that it requires estimating the entire joint distribution of unobservables, which is not necessarily identified, as I have only given conditions for identifying the marginals $\gamma_1 \mid X$ and $\gamma_2 \mid X$. Chen, Tamer, and Torgovitsky (2011) provide general conditions under which the distance between the sieve MLE and the identified set converges to zero. I use this result to show that estimates of some point identified features of the entire parameter converge to the unique true value. Specifically, I prove consistency of the sieve MLE of the marginal distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X$, allowing for the possibility that the full joint distribution of the unobservables is not point identified. This consistency result applies after using either theorem 1 or theorem 2 to achieve identification.

The objects of interest here are the distribution functions, their densities, or functionals of these distributions such as the mean and variance. Much of the existing work in nonparametrics on rates of convergence and asymptotic distribution theory has focused on finite dimensional parameters in models with an infinite dimensional nuisance parameter, and there is much to be done when the infinite dimensional parameter itself is of interest. Although I leave a detailed analysis of inference in this sieve MLE with partial identification setup to future work, one possibility is to apply the weighted bootstrap procedure of Chen et al. (2011).

4.1 Sieve maximum likelihood

I first give a brief description of the sieve MLE; additional details are in section 4.3. We observe the joint distribution of $(Y, Z, X) = (Y_1, Y_2, Z_1, Z_2, X)$. To reduce the dimension of the estimation

problem, assume $(\beta_1, \beta_2, \delta_1, \delta_2)$ are constant coefficients; the main consistency result can be generalized to allow these coefficients to be random as well. Assume that $(U_1, U_2, \gamma_1, \gamma_2) \mid X$ has a density with respect to the Lebesgue measure. The likelihood of Y given (Z, X) is

$$\begin{aligned} &f_{Y|Z,X}(y \mid z, x) \\ &= \int f_{U|X,\Gamma}(y_1 - \gamma_1 y_2 - \beta_1 z_1 - \delta_1' x, y_2 - \gamma_2 y_1 - \beta_2 z_2 - \delta_2' x \mid x, \gamma_1, \gamma_2) |1 - \gamma_1 \gamma_2| \, dF_{\Gamma|X}(\gamma_1, \gamma_2 \mid x). \end{aligned}$$

Let N denote the sample size and n index the observations. The log conditional likelihood of a random sample $\{(y_n, z_n, x_n)\}_{n=1}^N$ from (Y, Z, X) is

$$\begin{aligned} L_N(\alpha) = \sum_{n=1}^N \log \iint_{\text{supp}(\gamma_1,\gamma_2)} &f_{U_1,U_2,\gamma_1,\gamma_2|X}(y_{1n} - \gamma_1 y_{2n} - \beta_1 z_{1n} - \delta_1' x_n, \\ &y_{2n} - \gamma_2 y_{1n} - \beta_2 z_{2n} - \delta_2' x_n, \gamma_1, \gamma_2 \mid x_n) \cdot |1 - \gamma_1 \gamma_2| \, d\gamma_1 d\gamma_2. \end{aligned}$$

The unknown parameters are $\alpha \equiv (\beta_1, \beta_2, \delta_1, \delta_2, f_{U_1,U_2,\gamma_1,\gamma_2|X})$. Let $\mathscr{A}$ denote the parameter space. The maximum likelihood estimator solves

$$\sup_{\alpha \in \mathscr{A}} L_N(\alpha).$$

This estimator is usually infeasible since it requires optimization over an infinite-dimensional parameter space. To obtain a feasible version of this estimator, we can replace the infinite dimensional space $\mathscr{A}$ with a finite dimensional approximation $\mathscr{A}_N$. This approach is called the method of sieves, and $\mathscr{A}_N$ are called sieve spaces. An estimator $\hat{\alpha}_N$ which solves

$$\sup_{\alpha \in \mathscr{A}_N} L_N(\alpha) \tag{5}$$

is called a sieve maximum likelihood estimator. Let $\alpha_0 \in \mathscr{A}$ denote the true parameter value. Most consistency results for sieve estimators, such as theorem 3.1 of Chen (2007) or theorem 4.2 of Bierens (2012), require α_0 to be point identified. Since α_0 might be partially identified, these results do not apply. Chen et al. (2011), however, provide a general consistency theorem for sieve extremum estimators with partially identified parameters. In section 4.3, I apply their results to show that a sieve MLE of the entire parameter vector is consistent in some sense. I use this result to show that the sieve MLE of the pdfs of $\gamma_1 \mid X$ and of $\gamma_2 \mid X$ are consistent in the sup-norm. I discuss implementation of the sieve estimator and provide Monte Carlo simulations in appendix 5.

4.2 Two-stage least squares

As discussed in section 2.2, nearly parallel lines can preclude mean-based identification approaches. In this case, the reduced form mean regressions $E(Y_1 \mid X, Z)$ and $E(Y_2 \mid X, Z)$ 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 2) 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 $\hat{\gamma}_1$ denote the 2SLS estimator of γ_1. Assuming the relevant means exist (see section 2.2), we have

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

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 $E(\gamma_1)$, 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 section 5, I compare this weighted average effect parameter to $E(\gamma_1)$ in several example simulations. The difference is largest when the true distribution of random coefficients is highly skewed. In this paper, I have given conditions under which we can go beyond this weighted average effect parameter and identify the entire marginal distribution of each random coefficient.

4.3 Consistency of a sieve MLE

In this section I provide general conditions for a sieve MLE defined by equation (5) to be consistent. These conditions are essentially just those of Chen et al. (2011), with additional assumptions on the parameter space specific to the estimation problem considered here.

The likelihood of Y given (Z, X) and the log conditional likelihood of a random sample of (Y, X, Z) were given in section 4.1. I first derive those results in detail.

Assumption E1. Conditional on X, $(U, \Gamma) = (U_1, U_2, \gamma_1, \gamma_2)$ has a density with respect to the Lebesgue measure.

The likelihood of Y given (Z, X) is

$$
\begin{aligned}
&f_{Y|Z,X}(y \mid z, x) \\
&= \int f_{Y|Z,X,\Gamma}(y \mid z, x, \gamma_1, \gamma_2)\, dF_{\Gamma|Z,X}(\gamma_1, \gamma_2 \mid z, x) \\
&= \int f_{U|Z,X,\Gamma}(u \mid z, x, \gamma_1, \gamma_2)|1 - \gamma_1\gamma_2|\, dF_{\Gamma|Z,X}(\gamma_1, \gamma_2 \mid z, x) \\
&= \int f_{U|Z,X,\Gamma}(y_1 - \gamma_1 y_2 - \beta_1 z_1 - \delta_1' x, y_2 - \gamma_2 y_1 - \beta_2 z_2 - \delta_2' x \mid z, x, \gamma_1, \gamma_2)|1 - \gamma_1\gamma_2|\, dF_{\Gamma|Z,X}(\gamma_1, \gamma_2 \mid z, x) \\
&= \int f_{U|X,\Gamma}(y_1 - \gamma_1 y_2 - \beta_1 z_1 - \delta_1' x, y_2 - \gamma_2 y_1 - \beta_2 z_2 - \delta_2' x \mid x, \gamma_1, \gamma_2)|1 - \gamma_1\gamma_2|\, dF_{\Gamma|X}(\gamma_1, \gamma_2 \mid x) \\
&= \iint_{\mathrm{supp}(\gamma_1,\gamma_2)} f_{U_1,U_2,\gamma_1,\gamma_2|X}(y_1 - \gamma_1 y_2 - \beta_1 z_1 - \delta_1' x, \\
&\qquad\qquad y_2 - \gamma_2 y_1 - \beta_2 z_2 - \delta_2' x, \gamma_1, \gamma_2 \mid x) \cdot |1 - \gamma_1\gamma_2|\, d\gamma_1\, d\gamma_2 \\
&\equiv p(y \mid z, x; \alpha),
\end{aligned}
$$

where $\alpha = (\beta_1, \beta_2, \delta_1, \delta_2, f_{U_1,U_2,\gamma_1,\gamma_2|X})$. The fourth line follows since $Z \perp\!\!\!\perp (U, \Gamma) \mid X$. The second line follows since, by a change of variables,

$$
f_{Y|Z,X,\Gamma}(y \mid z, x, \gamma_1, \gamma_2) = f_{U|Z,X,\Gamma}(u \mid z, x, \gamma_1, \gamma_2) \left| \frac{\partial U}{\partial Y} \right|
$$

where $|\partial U / \partial Y|$ is the determinant of the Jacobian of the transformation of Y into U,

$$
U = (I - \Gamma)Y - BZ = \begin{pmatrix} Y_1 - \gamma_1 Y_2 - \beta_1 Z_1 - \delta_1' X \\ Y_2 - \gamma_2 Y_1 - \beta_2 Z_2 - \delta_2' X \end{pmatrix},
$$

and hence $|\partial U / \partial Y| = |\det(I - \Gamma)| = |1 - \gamma_1\gamma_2|$.

The log conditional likelihood of a random sample $\{(y_n, z_n, x_n)\}_{n=1}^N$ from (Y, Z, X) is

$$
L_N(\alpha) = \sum_{n=1}^N \log p(y_n \mid z_n, x_n; \alpha).
$$

For real-valued functions with domain $\mathcal{D} \subseteq \mathbb{R}^{d_x}$, denote the differential operator by

$$
\nabla^\lambda = \frac{\partial^{|\lambda|}}{\partial x_1^{\lambda_1} \cdots \partial x_{d_x}^{\lambda_{d_x}}} = \frac{\partial^{\lambda_1}}{\partial x_1^{\lambda_1}} \cdots \frac{\partial^{\lambda_{d_x}}}{\partial x_{d_x}^{\lambda_{d_x}}},
$$

where $\lambda = (\lambda_1, \ldots, \lambda_{d_x})$ is a multi-index, a d_x-tuple of non-negative integers, and $|\lambda| = \lambda_1 + \cdots + \lambda_{d_x}$. For a positive integer m, let $\mathscr{C}^m(\mathcal{D})$ denote the set of functions $f : \mathcal{D} \to \mathbb{R}$ such that $\nabla^\lambda f$ exists

and is continuous for all $|\lambda| \leq m$. For functions $f \in \mathscr{C}^m(\mathcal{D})$, define the sup norm

$$\|f\|_\infty \equiv \sup_{x \in \mathcal{D}} |f(x)|,$$

the Hölder norm

$$\|f\|_\Lambda \equiv \max_{|\lambda| \leq m} \|\nabla^\lambda f\|_\infty + \max_{|\lambda|=m} \sup_{x \neq x'} \frac{|\nabla^\lambda f(x) - \nabla^\lambda f(x')|}{(\|x - x'\|_e)^\gamma},$$

where $0 < \gamma \leq 1$ and $\|\cdot\|_e$ is the Euclidean norm on $\mathbb{R}^d$, and define the weighted norms

$$\|f\|_s \equiv \|f(\cdot)\omega_s(\cdot)\|_\Lambda$$
$$\|f\|_c \equiv \|f(\cdot)\omega_c(\cdot)\|_\infty,$$

where $\omega_s : \mathcal{D} \to \mathbb{R}$ and $\omega_c : \mathcal{D} \to \mathbb{R}$ are weighting functions. Let $\omega_s(x) = (1 + \|x\|_e^2)^{\zeta_s}$ and $\omega_c(x) = (1 + \|x\|_e^2)^{\zeta_c}$ where $\zeta_s > \zeta_c > d_x/2$. I will assume that the parameter space is bounded under $\|\cdot\|_s$, which yields compactness under $\|\cdot\|_c$. I then prove one-sided Hausdorff consistency of the overall sieve MLE under $\|\cdot\|_c$. I use this result to prove sup-norm consistency for the marginal distributions of random coefficients.

Assumption E2 (Parameter space). Let $\mathcal{X} \subseteq \mathbb{R}^K$ denote the support of X. Let $(U_1, U_2, \gamma_1, \gamma_2) \mid X = x$ have support $\mathcal{U} \times \mathcal{G} \subseteq \mathbb{R}^4$ for all $x \in \mathcal{X}$. Define $\mathscr{A} = \mathcal{B} \times \mathcal{D} \times \mathscr{F}$, where these parameter spaces satisfy the following:

1. $\mathcal{X}$ is a compact, nonempty subset of the Euclidean space $(\mathbb{R}^K, \|\cdot\|_e)$.
2. $\mathcal{B}$ is a compact, nonempty subset of the Euclidean space $(\mathbb{R}^2, \|\cdot\|_e)$. $0 \notin \mathcal{B}$.
3. $\mathcal{D}$ is a compact, nonempty subset of the Euclidean space $(\mathbb{R}^K, \|\cdot\|_e)$.
4. $\mathscr{F}$ is a $\|\cdot\|_c$-closed subset of the Hölder ball $\{f \in \mathscr{C}^m(\mathcal{U} \times \mathcal{G} \times \mathcal{X}) : \|f\|_s \leq B_0\}$, where m is a strictly positive integer, $\zeta_s > \zeta_c > (4 + K)/2$, $B_0 < \infty$ is a known constant, and such that for all $f \in \mathscr{F}$,
   (a) $f(u_1, u_2, \gamma_1, \gamma_2 \mid x) \geq 0$ for all $(u_1, u_2, \gamma_1, \gamma_2) \in \mathcal{U} \times \mathcal{G}$ and all $x \in \mathcal{X}$, and
   (b) $\int_{\mathcal{U} \times \mathcal{G}} f(u_1, u_2, \gamma_1, \gamma_2 \mid x)\, du_1\, du_2\, d\gamma_1\, d\gamma_2 = 1$ for all $x \in \mathcal{X}$.
5. For all $f \in \mathscr{F}$, $f_{\gamma_1|X}$ and $f_{\gamma_2|X}$ are identified.
6. $Q : \mathscr{A} \to [0, \infty)$ defined by $Q(\alpha) = E_0[\log p(Y \mid Z, X; \alpha)]$ is $(\|\cdot\|_{\mathscr{A}}, |\cdot|)$-continuous on $\mathscr{A}$.

Assume the true parameter value α_0 is in $\mathscr{A}$. Let $\|\alpha\|_{\mathscr{A}} \equiv \|(\beta_1, \beta_2)\|_e + \|(\delta_1, \delta_2)\|_e + \|f_{U_1,U_2,\gamma_1,\gamma_2|X}\|_c$.

The parameter space $\mathscr{F}$ is defined as a subset of a closed ball under the weighted Hölder norm $\|\cdot\|_s$. Following Gallant and Nychka (1987) and recent papers based on their work such as Newey

and Powell (2003) and Santos (2012), I show consistency of the sieve MLE under the norm $\|\cdot\|_c$, which is particularly aided by the fact that the parameter space $\mathscr{F}$ is compact under this norm (although it is not compact under $\|\cdot\|_s$). Restricting the joint density of the unobservables to lie in a weighted Hölder ball places restrictions on the tails of these distributions. In particular, since $\zeta_s > 0$, the weight function $\omega_s(\cdot) = (1 + \|\cdot\|_e^2)^{\zeta_s}$ puts large weight on large values of its argument. Consequently, since the weighted Hölder norm is bounded, the weighted sup-norm is also bounded:

$$\sup_{u_1,u_2,\gamma_1,\gamma_2,x} |f(u_1,u_2,\gamma_1,\gamma_2 \mid x)|(1 + \|(u_1,u_2,\gamma_1,\gamma_2,x)\|_e^2)^{\zeta_s} < B_0$$

for any $f \in \mathscr{F}$. This implies an upper bound on the tails of the density functions $f(u_1,u_2,\gamma_1,\gamma_2 \mid x)$: they must decrease at least as fast as the weight increases. Since $\mathcal{X}$ is assumed to be compact, this tail restriction does not actually restrict the way in which x affects the function.[4] This tail restriction does, however, restrict the kinds of distributions of $(u_1,u_2,\gamma_1,\gamma_2)$ to have tails which are not too fat. I discuss one approach to enforcing these assumptions in practice, especially the identification assumption E2.5, in section 5.

Assumption E3 (Sieve spaces).

1. For each $k \geq 1$, $\mathscr{A}_k = \mathcal{B} \times \mathcal{D} \times \mathscr{F}_k$, $\mathscr{F}_k \neq \emptyset$, $\mathscr{F}_k \subseteq \mathscr{F}$, $\mathscr{F}_k \subseteq \mathscr{F}_{k+1}$, and $\dim(\mathscr{F}_k) < \infty$.
2. $\mathscr{F}_k$ is compact under $\|\cdot\|_c$.
3. $\cup_{k=1}^{\infty} \mathscr{F}_k$ is dense in $\mathscr{F}$ under $\|\cdot\|_c$: For any $f \in \mathscr{F}$, there is an element $\pi_k f \in \mathscr{F}_k$ such that $\|f - \pi_k f\|_c \to 0$ as $k \to \infty$.

I discuss examples of such sieve spaces in appendix 5.

Assumption E4 (Uniform convergence).

1. The data $\{(y_n, z_n, x_n)\}_{n=1}^N$ are a random sample of (Y,Z,X) from a unique density p_0.
2. For each N, $E_0(\sup_{\alpha \in \mathscr{A}_N} |\log p(Y \mid Z, X; \alpha)|)$ is bounded.
3. (Hölder condition) There is a finite $s > 0$ and a positive function $U_N(y,z,x)$ with $E_0[U_N(Y,Z,X)] < \infty$ such that

$$\sup_{\alpha,\alpha' \in \mathscr{A}_N : \|\alpha - \alpha'\|_{\mathscr{A}} \leq \delta} |\log p(y \mid z, x; \alpha) - \log p(y \mid z, x; \alpha')| \leq \delta^s U_N(y,z,x)$$

   for all z.
4. (Entropy condition) For all $\delta > 0$, the sieve spaces satisfy

$$H(\delta^{1/s}, \mathscr{A}_N, \|\cdot\|_{\mathscr{A}}) = o(N),$$

[4] We can allow regressors with unbounded support, but the norm must be adjusted appropriately to prevent the weighted Hölder ball assumption from being too restrictive.

where $H(\delta^{1/s}, \mathscr{A}_N, \|\cdot\|_{\mathscr{A}})$ denotes the log of the minimal number of $\delta^{1/s}$-radius balls (under the metric induced by $\|\cdot\|_{\mathscr{A}}$) that cover the space $\mathscr{A}_N$.

Theorem 3. Assume E1-E4. Let $\hat{\mathscr{A}_N}$ denote the set of solutions to the sieve maximum likelihood problem (5). Then for any $\hat{\alpha}_N \in \hat{\mathscr{A}_N}$, we have

$$\inf_{\alpha \in \mathscr{A}_I} \|\hat{\alpha}_N - \alpha\|_{\mathscr{A}} = o_p(1) \quad \text{and} \quad \|\hat{f}_{\gamma_i|X} - f_{\gamma_i|X}\|_\infty = o_p(1) \text{ for } i = 1, 2,$$

where

$$\hat{f}_{\gamma_1|X}(\gamma_1 \mid x) \equiv \int \hat{f}_{U_1,U_2,\gamma_1,\gamma_2|X}(u_1, u_2, \gamma_1, \gamma_2 \mid x)\, du_1\, du_2\, d\gamma_2$$

and likewise for $\hat{f}_{\gamma_2|X}$.

The proof is in appendix A.

5 Implementing the sieve estimator and Monte Carlo simulations

In this section I discuss one approach to implementing the sieve estimator, which I then use in Monte Carlo simulations to examine the estimator's finite sample performance. I make several substantive assumptions which help mitigate the curse of dimensionality and yet still allow some modeling flexibility.

5.1 Implementing the sieve estimator

Recall that the likelihood of a single observation Y given (Z, X) is

$$f_{Y|Z,X}(y \mid z, x) = \iint_{\text{supp}(\gamma_1,\gamma_2)} f_{U_1,U_2,\gamma_1,\gamma_2|X}(y_1 - \gamma_1 y_2 - \beta_1 z_1 - \delta_1' x, \\ y_2 - \gamma_2 y_1 - \beta_2 z_2 - \delta_2' x, \gamma_1, \gamma_2 \mid x) \cdot |1 - \gamma_1\gamma_2|\, d\gamma_1 d\gamma_2.$$

In principle, the theory of section 4.3 allows us to nonparametrically estimate $f_{U_1,U_2,\gamma_1,\gamma_2|X}$ and then use this estimate to project out the marginal distributions of f_{γ_1} and f_{γ_2}. This procedure involves estimating a $4 + \dim(X)$ dimensional function, and hence the estimator will likely have a very slow rate of convergence, resulting in a poor finite sample approximation with practical sample sizes. To lessen this problem, I make the following assumptions.

Assumption E5 (Dimension reduction).

1. (U_1, U_2) and (γ_1, γ_2) are independent, conditional on X.
2. X is independent of (U_1, U_2) and of (γ_1, γ_2).

3. The dependence between γ_1 and γ_2 can be modeled by a parametric copula. Specifically, the cdf of (γ_1, γ_2) is
$$F_{\gamma_1,\gamma_2}(t_1, t_2) = C(F_{\gamma_1}(t_1), F_{\gamma_2}(t_2); \rho_\gamma),$$
where $C(\cdot, \cdot; \rho_\gamma)$ is a known parametric copula with parameter $\rho_\gamma \in \mathbb{R}^{\dim(\rho_\gamma)}$.

4. γ_1 and γ_2 have the same marginal distribution: $F_{\gamma_1} = F_{\gamma_2}$.

5. (U_1, U_2) are bivariate normally distributed with zero means and identical marginal distributions (i.e. equal variances $\sigma_{U_1} = \sigma_{U_2}$).

E5.1 holds under constant coefficients, but it also allows for arbitrary distributions of (γ_1, γ_2). I make E5.1 to avoid having to choose a sieve space for functions with different domains for different arguments (e.g. the domain $\mathbb{R}^2 \times [0,1]^2$). In addition to our earlier assumption that $(\beta_1, \beta_2, \delta_1, \delta_2)$ are constant, E5.2 is the main dimension reduction assumption. It rules out heteroskedasticity, among other things. E5.3 allows for specific dependence patterns between γ_1 and γ_2 via the choice of the parametric copula function. Copulas are used to separately model marginal and joint distributions. See Nelson (2006) for further discussion of copulas and examples of parametric copulas (see page 116). For a related application combining parametric copulas and nonparametric marginal distributions using sieves, see Chen, Fan, and Tsyrennikov (2006). E5.5 also allows for dependence between U_1 and U_2. E5.4 and E5.5 both impose a symmetric marginal assumption. This assumption holds in the empirical application where the two equations correspond to people. Since the labels of person 1 versus person 2 are arbitrary, the marginal distributions of variables for each person are the same.

None of the assumptions in E5 are necessary for consistency; any of them can be relaxed as desired. For example, E5.2 can be relaxed to allow for specific forms of parametric heteroskedasticity. Likewise, we can perform all analysis conditional on values of discrete X's, thus allowing for X to affect the entire distribution of (γ_1, γ_2). A scale and location model can be used to allow for continuous X's to affect (γ_1, γ_2), although this is not straightforward since I assume the random coefficients have bounded support (see below). I make E5.5 in order to focus on estimating the distribution of random coefficients nonparametrically, but E5.5 can be relaxed along the lines of E5.3 and E5.4. That is, we can model the joint distribution of (U_1, U_2) using a parametric copula, impose symmetric marginals, and then estimate the common marginal distribution via sieves. Furthermore, we can relax the parametric copula assumption by estimating the copula itself using sieves (see Sancetta and Satchell 2004).

In addition to the dimension reduction assumptions, I restrict the support of the random coefficients.

Assumption E6 (Support restriction). F_γ has support $[0, \theta]$ where $\theta > 0$ is a known constant.

If $\theta < 1$, then this support restriction, combined with E5.5 (bivariate normal additive unobservables) and the instrument requirements, relevance (A2), independence (A3), and continuous

variation (A4′), ensure that F_γ is identified, by proposition 2 and theorem 2. $\theta \geq 1$ is allowed, but in this case identification is only guaranteed if one is willing to make the reduced form moment assumptions A5 to apply theorem 2 or if one has a full support instrument in order to apply theorem 1. E6 may be relaxed to allow negative values, but the same caveats to including 1 in the support apply to including -1 in the support.

Given the above assumptions, the likelihood of Y given (Z, X) simplifies to

$$f_{Y|Z,X}(y \mid z, x) = \iint_{[0,\theta]^2} \phi_{\sigma_u,\rho_u}(y_1 - \gamma_1 y_2 - \beta_1 z_1 - \delta_1' x, y_2 - \gamma_2 y_1 - \beta_2 z_2 - \delta_2' x) \cdot c(F_\gamma(\gamma_1), F_\gamma(\gamma_2); \rho_\gamma) f_\gamma(\gamma_1) f_\gamma(\gamma_2) |1 - \gamma_1\gamma_2| \, d\gamma_1 d\gamma_2,$$

where ϕ_{σ_u,ρ_u} is the symmetric bivariate normal pdf with variance σ_u^2 and correlation ρ_u, $c(\cdot,\cdot;\rho_\gamma)$ is the density for the copula cdf $C(\cdot,\cdot;\rho_\gamma)$, and F_γ and f_γ are the cdf and pdf, respectively, of the common marginal distribution of γ_1 and γ_2. This common marginal distribution is the only nonparametric component remaining in the likelihood. In practice, we must choose a particular sieve space to approximate this distribution. I approximate the density f_γ by

$$f_{\gamma;J_N}(s) = \frac{[\text{SPL}(s, M, J_N)]^2}{\int_0^\theta [\text{SPL}(v, M, J_N)]^2 \, dv},$$

where

$$\text{SPL}(s, M, J_N) = \sum_{m=0}^{M} a_m s^m + \sum_{j=1}^{J_N} b_j [\max\{s - t_j, 0\}]^M$$

for sieve coefficients $a_m, b_j \in \mathbb{R}$, and M and J_N are positive integers, and $0 = t_0 < t_1 < \cdots < t_{J_N} < t_{J_N+1} = \theta$ is a partition of $[0, \theta]$. I assume the knots t_j are equally spaced. $\text{SPL}(s, M, J_N)$ is a polynomial spline of order M with J_N knots. See Chen (2007) pages 5569-5580 for a discussion of other sieve spaces.

This form of the density approximation ensures that the estimated density is non-negative and integrates to one. The sieve estimator requires computing two integrals: the denominator of $f_{\gamma;J_N}$ and the integral over the random coefficients. Any numerical integration method can be used. I use Gauss-Legendre quadrature for the denominator approximation and I use Halton draws to integrate over (γ_1, γ_2). Evaluating the copula density requires estimating the cdf F_γ. To do this, I integrate the estimated density: $F_{\gamma;J_N}(s) = \int_0^s f_{\gamma;J_N}(v) \, dv$.

In addition to the sieve space choice, we must choose a parametric copula function. I use the Gaussian copula of dimension 2:

$$C(u_1, u_2; \rho_\gamma) = \Phi_{\rho_\gamma}(\Phi^{-1}(u_1), \Phi^{-1}(u_2)),$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cdf, and $\Phi_{\rho_\gamma}(\cdot,\cdot)$ is the cdf for the bivariate normal with unit variances and correlation parameter $\rho_\gamma \in [-1, 1]$. This copula allows for γ_1 and

γ_2 to be independent if $\rho_\gamma = 0$, positively related if $\rho_\gamma > 0$, and negatively related if $\rho_\gamma < 0$.

Finally, note that choosing a parametric specification for f_γ, such as the truncated normal distribution or the beta distribution, leads to a fully parametric maximum likelihood estimator. Both the fully parametric estimator and the sieve estimator are analogous to the classic full information maximum likelihood (FIML) estimator of a simultaneous equations system, which assumes that the coefficients are unknown constants and the additive errors are jointly normally distributed.

5.2 Monte Carlo simulations

To examine the sieve estimator's finite sample performance, I run several Monte Carlo simulations. The conditions of both theorems 1 and 2 hold in all simulations so that either result could be used to ensure identification. I consider three different data generating processes. All dgps are identical except for the common marginal distribution f_γ, which 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.
3. f_γ is a truncated normal with pre-truncation mean 0 and standard deviation 0.2.

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 $[0, \theta]$. For each dgp I consider the sample sizes $N = 800$ and $N = 400$, which are approximately the size of the full and restricted samples, respectively, in my empirical application. All dgps have γ_1 independent of γ_2. All dgps use the same distribution of additive errors and of the covariates, and the same constant covariate coefficients. (U_1, U_2) are bivariate normal with $\mu_u = 0$, $\sigma_u = 1$, and $\rho_u = 0$. There are four covariates. Two covariates, Z_1 and Z_2, have a $\mathcal{N}(0, 3)$ distribution, 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. Two covariates, X_1 and X_2, have a $\mathcal{N}(5, 1)$ distribution. The own coefficient on X_1 is 6 and the friend coefficient is 3. Both the own and friend coefficients on X_2 are 0. The constant term is -10. $\theta = 0.95$, which ensures that the common marginal distribution f_γ is identified. The true structural system with these parameter values is

$$Y_1 = -10 + \gamma_1 Y_2 + (5Z_{11} + 0Z_{12}) + (0Z_{21} + 0Z_{22}) + (6X_{11} + 3X_{12}) + (0X_{21} + 0X_{22}) + U_1$$
$$Y_2 = -10 + \gamma_2 Y_1 + (5Z_{12} + 0Z_{11}) + (0Z_{22} + 0Z_{21}) + (6X_{12} + 3X_{11}) + (0X_{22} + 0X_{21}) + U_2.$$

Although the second Z and the second X covariate both have zero own and friend coefficients, they are treated differently by the estimator since the exclusion restriction (zero friend coefficient) is imposed for Z but not for X.

For each dgp, I compute several statistics. First, I compute the bias and median 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. I use $S = 250$ simulations, which yields simulation standard errors small enough to make statistically significant comparisons. The estimated median bias is the median of $\hat{\kappa}_s - \kappa$ over all $s = 1, \ldots, S$. I compute these statistics for the sieve estimator of the random coefficients' mean:

$$\widehat{E(\gamma)} = \int_0^\theta x \cdot \hat{f}_{\gamma;J_N}(x)\, dx,$$

as well as for the 2SLS estimator of the endogenous variable coefficient, viewed as an estimator of $E(\gamma)$. I also compute these statistics for the sieve and 2SLS estimators of β_1, the constant coefficient on the instrument Z_1. Finally, I compute the mean and median integrated squared error of the sieve density estimator $\hat{f}_{\gamma;J_N}$ of f_γ. For a fixed simulation s, the ISE is

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

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

In the simulations and the empirical application, I choose $J_N = 14$ for the smaller sample, $J_N = 22$ for the larger sample, and let $M = 3$ for both samples. Choosing size of the sieve space leads to a tradeoff between bias and variance of the sieve estimator: The larger the space, the smaller the bias and the higher the variance. I use a relatively large number of knots to accommodate all three dgps.

Figure 2 shows example plots of $\hat{f}_{\gamma;J_N}$ versus the true density. Table 1 shows the estimated bias and MISE for each of the three dgps and the two sample sizes.

Table 1: Monte Carlo results: Means

	Bias in $\widehat{E(\gamma)}$		Bias in $\hat{\beta}_1$		MISE
	Sieve	2SLS	Sieve	2SLS	
Indep. trunc. normal(0.4,0.05)	$E(\gamma) = 0.38$		$\beta_1 = 5$		
$N = 400$	0.0016	0.0010	-0.0011	0.0033	0.1044
	[0.0054]	[0.0093]	[0.0384]	[0.0564]	[0.0733]
$N = 800$	-0.0016	0.0010	0.0030	0.0011	0.1442
	[0.0071]	[0.0066]	[0.0311]	[0.0395]	[0.2124]
Indep. Beta(6,3)	$E(\gamma) = 0.63$		$\beta_1 = 5$		
$N = 400$	-0.0052	0.0236	0.0046	0.0255	0.0480
	[0.0296]	[0.0318]	[0.2262]	[0.3258]	[0.1408]
$N = 800$	-0.0089	0.0237	0.0462	0.0126	0.0319
	[0.0227]	[0.0221]	[0.1590]	[0.2161]	[0.0798]
Indep. trunc. normal(0,0.2)	$E(\gamma) = 0.15$		$\beta_1 = 5$		
$N = 400$	0.0516	0.1166	0.0221	0.0251	0.3281
	[0.0399]	[0.0376]	[0.0819]	[0.2005]	[0.8325]
$N = 800$	0.0424	0.1167	0.0252	0.0126	0.1801
	[0.0364]	[0.0262]	[0.0715]	[0.1377]	[0.3655]

Standard deviations in brackets; square these and add to the squared bias to obtain MSE.

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 sieve and the 2SLS estimator estimate $E(\gamma)$ well. The second dgp is slightly asymmetric and more spread out. In this case, both estimators do worse than in the first dgp, but the 2SLS estimator's bias in estimating $E(\gamma)$ is more than twice as large as the sieve estimator. The third dgp is highly skewed and both the sieve and 2SLS estimators of $E(\gamma)$ perform worse than the first two dgps. In this case, the 2SLS estimator is again worse than the sieve, with a bias more than twice as large as the sieve estimator. For the first dgp, the sieve estimator provides a good fit even at $N = 400$, so an increase in sample size does not change much. For the other two dgps, doubling the sample size reduces the MISE, as well as the root-MSE of the scalar parameter estimates. The bias in 2SLS does not change with sample size, as expected since 2SLS is inconsistent for $E(\gamma)$—instead it is consistent for a weighted average of γ.

Table 2 shows the median biases and the median ISE. The overall patterns from table 1 hold here as well. The more skewed distributions make $E(\gamma)$ harder to estimate. Now, however, the

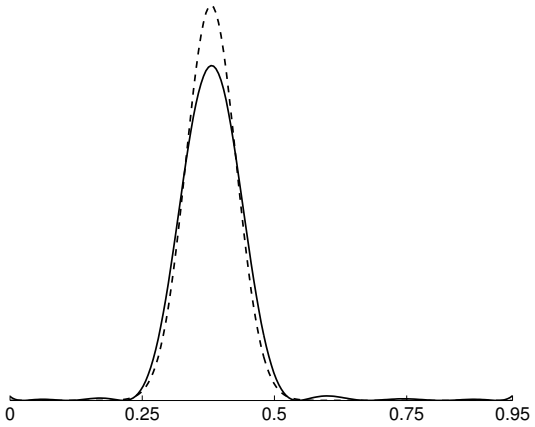
sieve estimator of $E(\gamma)$ has median bias from anywhere from 3.5 times to 6.5 times smaller than the 2SLS estimator in the second and third dgps.

Table 2: Monte Carlo results: Medians

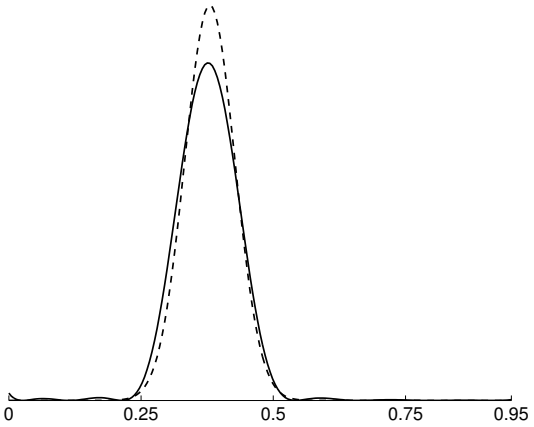
	Med. bias in $\widehat{E(\gamma)}$		Med. bias in $\hat{\beta}_1$		Median ISE
	Sieve	2SLS	Sieve	2SLS	
Indep. trunc. normal(0.4,0.05)	$E(\gamma) = 0.38$		$\beta_1 = 5$		
$N = 400$	0.0023	0.0018	-0.0017	0.0038	0.0937
$N = 800$	-0.0013	0.0007	0.0035	0.0060	0.0994
Indep. Beta(6,3)	$E(\gamma) = 0.63$		$\beta_1 = 5$		
$N = 400$	-0.0041	0.0265	0.0121	-0.0013	0.0205
$N = 800$	-0.0067	0.0236	0.0559	0.0162	0.0169
Indep. trunc. normal(0,0.2)	$E(\gamma) = 0.15$		$\beta_1 = 5$		
$N = 400$	0.0303	0.1191	0.0132	0.0190	0.0645
$N = 800$	0.0208	0.1166	0.0245	0.0119	0.0170

The third dgp resembles what we might expect to see in the empirical application, and this is precisely when the bias in 2SLS is largest. This bias, about 0.12, is economically large.

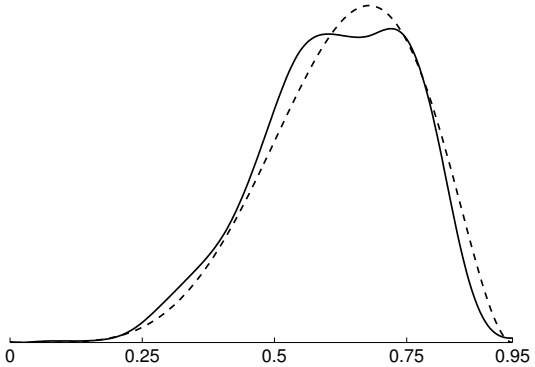
Overall, the simulation results suggest that the sieve estimator performs well with practical sample sizes. Skewed distributions, which we might expect in practice, result in larger biases for both the sieve and 2SLS estimators, but the sieve estimator significantly outperforms 2SLS.



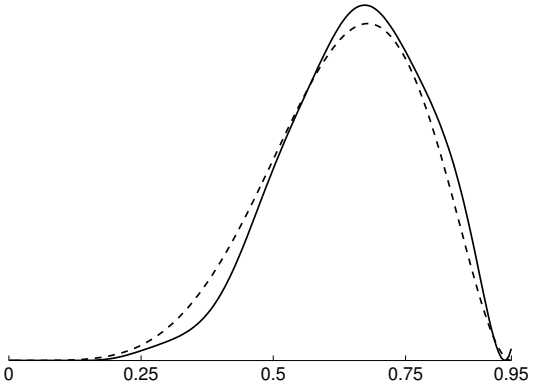
(a) Indep. trunc. normal(0.4,0.05), $N = 400$



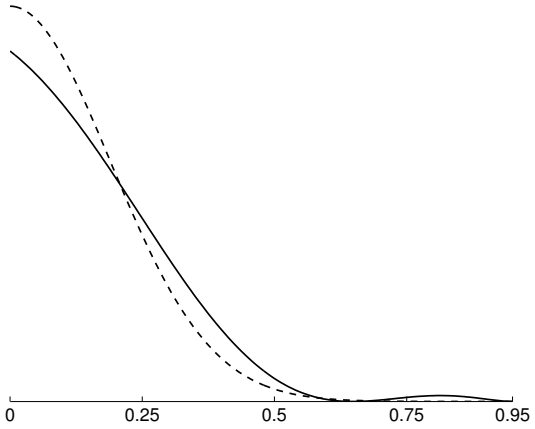
(b) Indep. trunc. normal(0.4,0.05), $N = 800$



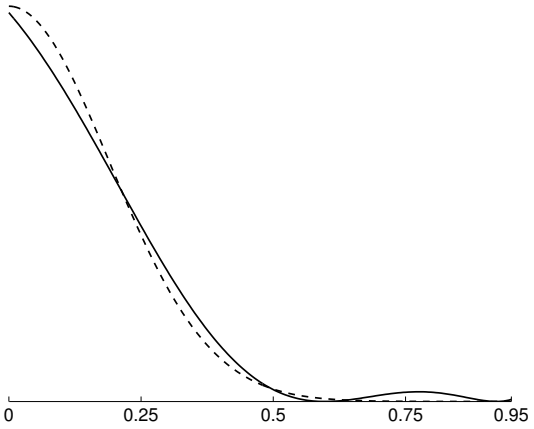
(c) Indep. Beta(6,3), $N = 400$



(d) Indep. Beta(6,3), $N = 800$



(e) Indep. trunc. normal(0,0.2), $N = 400$



(f) Indep. trunc. normal(0,0.2), $N = 800$

Figure 2: Sieve 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.

6 The social determinants of obesity

In this section, I use the methods developed in this paper to explore the social determinants of obesity. A large and controversial literature on this topic has developed since Christakis and Fowler (2007) concluded that "obesity appears to spread through social ties". I construct pairs of best friends using the Add Health dataset (Harris, Halpern, Whitsel, Hussey, Tabor, Entzel, and Udry 2009). I then apply the sieve estimator to nonparametrically estimate the distribution of random coefficients γ_1 and γ_2 in the simultaneous equations model (1), where outcomes are changes in weight between two time periods and the instruments are changes in height between the same two periods. This approach yields estimates of the average social effect while allowing that not all people affect their best friend equally.

6.1 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 waves of data collection. I use data from the wave 1 in-home survey, administered between April and December 1995, and the wave 2 in-home survey, administered between April and August 1996. In both surveys, 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, including obesity (Cohen-Cole and Fletcher 2008, Fowler and Christakis 2008, Halliday and Kwak 2009, Renna, Grafova, and Thakur 2008, Trogdon, Nonnemaker, and Pais 2008). 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.

20,745 students are in the wave 1 data. 14,738 students are in the wave 2 data. From the sample of students remaining after wave 2, I construct 755 same-sex pairs of students—1510 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. This gives 476 pairs. I call this the restricted sample. 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. Continuing in this manner yields 279 additional pairs. Note that no student is included more than once. Although students were asked to name friends during both wave 1 and wave 2, I only use friendship data from the in-home wave 1 survey. I do not consider changes in friendship.

6.2 Empirical results

Research on obesity focuses on unhealthy weight change. To do this, most studies choose a measure of 'fatness' as the outcome variable, such as body mass index, which is weight in kilograms divided by squared height in meters. Instead of first scaling weight by height, I use weight directly as the outcome and then I condition on height. Specifically, I take the outcome of interest, Y_1 and Y_2 in model (1), to be the change in weight between the two waves for each student in a pair. I include both students' wave 1 heights as control variables in each equation. I also include both students' wave 1 weights as control variables. This allows for friendship formation based on weight, which would lead to students' baseline weights being correlated under positive assortative matching.

I use change in height between two waves as the instrument. In order to apply the identification result theorem 2 of section 3.2, any instrument must satisfy four conditions: (1) relevance, (2) exclusion, (3) independence, and (4) have continuous variation. Relevance is satisfied since weight increases are physically caused by height increases, holding all else constant. The exclusion restriction states that a change in one student's height cannot directly cause a change in the other student's weight, which seems plausible. Independence is satisfied if we assume height increases are caused by random growth spurts. Finally, height is a continuous variable. In the data, however, change in height is measured in inches and takes 31 distinct values. I discuss this discreteness further below.

Table 3: Summary statistics

	count	median	mean	sd	min	max
Weight change	1478	2.3	2.27	4.95	-23	31
Weight	1492	59	62.30	14.38	33	136
Height change	1488	0	0.02	0.03	0	.3
Height	1502	1.7	1.68	0.10	1.4	2
Smoking change	1492	0	1.16	7.96	-30	30
Health status change	1510	0	-0.01	0.84	-3	3
Age	1510	16	15.37	1.47	12	19

Weight is measured in kilograms, height in meters. All baseline variables (such as age) are measured at wave 1. Change variables are the difference between the wave 2 and wave 1 values. Count is number of people with non-missing values.

I include three additional control variables: smoking change, health status change, and age. In both waves, students were asked, "During the past 30 days, on how many days did you smoke cigarettes?" Smoking change is the difference in students' answers from wave 1 to wave 2. In both waves, students were asked to rank their general health from excellent (1) to poor (5). Health status change is the difference in students' answers from wave 1 to wave 2. Finally, I include students' age at wave 1. For all three of these variables, I assume that only a students' own value of the

variable affects their outcomes—their friends' smoking change, health status change, and age do not affect their weight change directly. These exclusion restrictions mean no exogenous social effects are included beyond one's friend's baseline weight and height. Consequently, variation in these covariates will also be used to aid estimation of endogenous social effects, which helps alleviate the discreteness in change in height. Table 3 shows summary statistics for all variables and all observations in the full sample. After dropping observations with missing covariate values, the number of friend pairs in the full sample is 691 and the number of friend pairs in the restricted sample is 424.

Table 4: Estimates of endogenous social interaction effect

	Full sample	Restricted sample
3SLS	.3216	.3810
Sieve $\widehat{E}(\gamma)$	.2357	.3765
Sieve $\widehat{\text{var}}(\gamma)$	.1484	.1518
Observations	691	424
Controls?	Yes	Yes

Observations are pairs of best friends. Weight changes in each friend are the endogenous variables. Weight is measured in kilograms, height in meters. Controls include own and friend's baseline height and weight (measured at wave 1), and own height change, smoking change, health change, and age (measured at wave 1). Change variables are the difference between the wave 2 and wave 1 values. Restricted sample consists only of pairs of people who named each other as their first best friend. Observations with any missing data are dropped. See body text for details of estimation.

Table 4 shows the estimation results. First, 3SLS provides estimates of system (1) under the assumption that all coefficients are constant, and under the restriction that the coefficients on each equation are equal ($\gamma_1 = \gamma_2$, $\beta_1 = \beta_2$, $\delta_1 = \delta_2$). The latter restriction holds since the labels of friend 1 versus friend 2 are arbitrary. Thus, we obtain a single point estimate of γ for each sample, shown in the first row of the table. The 3SLS point estimate of the social interaction effect for the full sample implies that a one kilogram increase in your friend's weight increases your own weight by 0.32 kilograms (the same effect size holds for pounds). This point estimate increases to 0.38 when considering only pairs of mutual first best friends. Both point estimates are statistically significant at the 5% level (p-value is 0.011 for the full sample, 0.009 for the restricted sample).

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, as shown in the simulation evidence in section 5, these estimates can be quite different from

the actual average coefficient. The sieve MLE estimator, on the other hand, provides a consistent estimator of the average random coefficient, as well as their distribution.

Estimates obtained from the sieve MLE are shown in the second and third row of table 4. As mentioned above, the labels of friend 1 versus friend 2 are arbitrary, so I assume the distribution of γ_1 equals the distribution of γ_2. Assume this distribution has support on $[0, 1]$ (using the support $[0, 0.95]$ as in the simulations makes little difference). Other assumptions and details of implementing the sieve estimator are as in section 5. I focus on estimates of two functionals of this distribution: the mean and the variance. The mean is easily comparable to 3SLS estimates while the variance provides a measure of the amount of heterogeneity. Functionals which involve averaging, like the mean and variance, are also usually estimated much more precisely than entire functions. This is particularly relevant here given the small sample sizes and that the instrument's variation is both discrete and quite small, since most students do not grow much over a single year (see table 3).

For the full sample, the sieve estimate of the average endogenous social interaction effect is 0.24, compared to the larger 3SLS estimate of 0.32. Moreover, the estimated variance in social interaction effects is 0.15, which is quite large. For the restricted sample, the sieve mean estimate and the 3SLS estimate are about the same, 0.38. The estimated variance is also approximately the same as in the full sample, 0.15. Overall, these results suggest that for many students, social influence matters for changes in weight, which is consistent with the existing empirical literature. In both samples, the sieve estimated mean is weakly smaller than 3SLS, suggesting that findings of social interaction effects based on 2SLS or 3SLS may overstate potential multiplier effects of policy interventions. Conversely, the sieve estimates also reveal substantial variation in social interaction effects, which suggests that there are some people who are highly susceptible to social interactions. The approach here has been to estimate unconditional means and variances. By instead estimating distributions of social interaction effects conditional on covariates, we can potentially explain some of the observed variation in these endogenous effects and learn which covariate combinations lead to large average effects. Interventions which target people with these covariates may have larger benefits than previously thought.

7 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 the social determinants of obesity and found evidence of significant heterogeneity as well as mean coefficient estimates equal to or smaller than the usual point estimates.

Several issues remain for future research. The sieve MLE of the two-equation model requires estimating several nuisance distributions. An alternative approach would be welcome. For any new approach, as well as for the proposed sieve estimator, inference must also be considered. Finally, it remains to be seen whether additional functionals of the full joint distribution of unobservables, such as the joint distribution of endogenous variable random coefficients, can be identified, and to what extent nonparametric identification can be achieved in the many-equation case.

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.

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

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*.

Bierens, H. (2012): "Consistency and asymptotic normality of sieve estimators under weak and verifiable conditions," *Working paper*.

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*.

Blundell, R. and R. L. Matzkin (2010): "Conditions for the existence of control functions in nonseparable simultaneous equations models," *Working paper*.

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

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

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, Forthcoming.

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.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6B, 5549–5632.

Chen, X., Y. Fan, and V. Tsyrennikov (2006): "Efficient estimation of semiparametric multivariate copula models," *Journal of the American Statistical Association*, 101, 1228–1240.

Chen, X., E. Tamer, and A. Torgovitsky (2011): "Sensitivity analysis in semiparametric likelihood models," *Working paper*.

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.

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.

Dunford, N. and J. T. Schwartz (1958): *Linear operators, part 1: general theory*, Interscience Publishers.

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

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.

Fowler, J. and N. Christakis (2008): "Estimating peer effects on health in social networks: A response to Cohen-Cole and Fletcher; Trogdon, Nonnemaker, Pais," *Journal of Health Economics*, 27, 1400.

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

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

Gallant, A. and D. Nychka (1987): "Semi-nonparametric maximum likelihood estimation," *Econometrica*, 55, 363–390.

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.

Halliday, T. and S. Kwak (2009): "Weight gain in adolescents and their peers," *Economics & Human Biology*, 7, 181–190.

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*.

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. (2013): "Identification in general triangular systems," *Working paper*.

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

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

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

Manski, C. F. (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.

——— (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.

Nelson, R. B. (2006): *An introduction to copulas*, Springer, second ed.

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

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.

Renna, F., I. Grafova, and N. Thakur (2008): "The effect of friends on adolescent body weight," *Economics & Human Biology*, 6, 377–387.

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

Rubin, H. (1950): "Note on random coefficients," in *Statistical Inference 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. (2001): "Peer effects with random assignment: results for dartmouth roommates," *The Quarterly Journal of Economics*, 116, 681–704.

Sancetta, A. and S. Satchell (2004): "The Bernstein copula and its applications to modeling and approximations of multivariate distributions," *Econometric Theory*, 20, 535–562.

Santos, A. (2012): "Inference in nonparametric instrumental variables with partial identification," *Econometrica*, 80, 213–275.

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. (2012): "Identification of nonseparable models with general instruments," *Working paper*.

Trogdon, J., J. Nonnemaker, and J. Pais (2008): "Peer effects in adolescent overweight," *Journal of Health Economics*, 27, 1388–1399.

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.

A Proofs

Proof of theorem 1. The proof has two 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 $\gamma_2 \mid X$.

1. Fix an $x \in \text{supp}(X)$. For any $z \in \text{supp}(Z)$, 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 and A4, we can apply lemma 1 to show that 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 $X = x$ is identified, for each $(t_1, t_2) \in \mathbb{R}^2$. In particular, note that the joint distribution of $(\pi_{11}, \pi_{12}, \pi_{13})$ given $X = x$ is identified by choosing $(t_1, t_2) = (1, 0)$, and the joint distribution of $(\pi_{21}, \pi_{22}, \pi_{23})$ is identified by choosing $(t_1, t_2) = (0, 1)$. These distributions will be used for steps (3) and (4).

2. Consider the term $t_1\pi_{11} + t_2\pi_{21}$. 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_{11}, π_{21}) is

$$\phi_{\pi_{11},\pi_{22}}(t_1, t_2) = E[\exp(i(t_1\pi_{11} + t_2\pi_{22}))].$$

The right hand side is identified for each $(t_1, t_2) \in \mathbb{R}^2$ and hence the characteristic function $\phi_{\pi_{11},\pi_{22}}$ is identified. Thus the joint distribution of (π_{11}, π_{21}) is identified, given $X = x$. Likewise, the joint distribution of (π_{12}, π_{22}) is identified, given $X = x$, and the joint distribution of (π_{13}, π_{23}) 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 3 implies that $\gamma_1 \mid X$ is identified.[5] 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 3 implies that $\gamma_2 \mid X$ is identified.

□

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

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

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

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

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 $t_n \neq 0$,

[5] 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 3 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 $\gamma_2 = \pi_{22}/\pi_{12}$.

$t_n \to 0$ 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. □

Lemma 3. 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 3. The distribution of X is identified directly from the observed marginal distribution of (YX, X). Next, we have

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

Thus, for $x > 0$,

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

and, for $x < 0$,

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

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

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

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. Identification of the joint distribution of $(\gamma_1\beta_2, \beta_2)$ follows from the proof of theorem 1. The result then follows by applying lemma 3. □

Proof of theorem 2. The proof strategy follows the same two steps as in the proof of theorem 1.

1. Use lemma 2 instead of lemma 1 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, A4′, and A5.

2. As in theorem 1.

□

Proof of lemma 2.

1. *Preliminary definitions and notation.* Let L be an arbitrary closed subspace of $\mathbb{R}^{K+1}$. Let $\text{proj}_L : \mathbb{R}^{K+1} \to L$ 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 $\ell(\lambda, \lambda z_1, \ldots, \lambda z_K)$. 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

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

(Note that these sets are not necessarily equal since $F_{\ell(1,z)} = \tilde{F}_{\ell(1,z)}$ 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 $\hat{z} = (1, z) \in R$ we have

$$\begin{aligned} \int (\hat{z}'y)^n dF(y) &= \int (t)^n dF_{\ell(1,z)}(t) \\ &= \int (t)^n d\tilde{F}_{\ell(1,z)}(t) \\ &= \int (\hat{z}'y)^n d\tilde{F}(y). \end{aligned}$$

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 4. (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: supp(Z) contains an open ball and thus R contains a cone in $\mathbb{R}^{K+1}$ (see figure 3), which has positive Lebesgue measure.

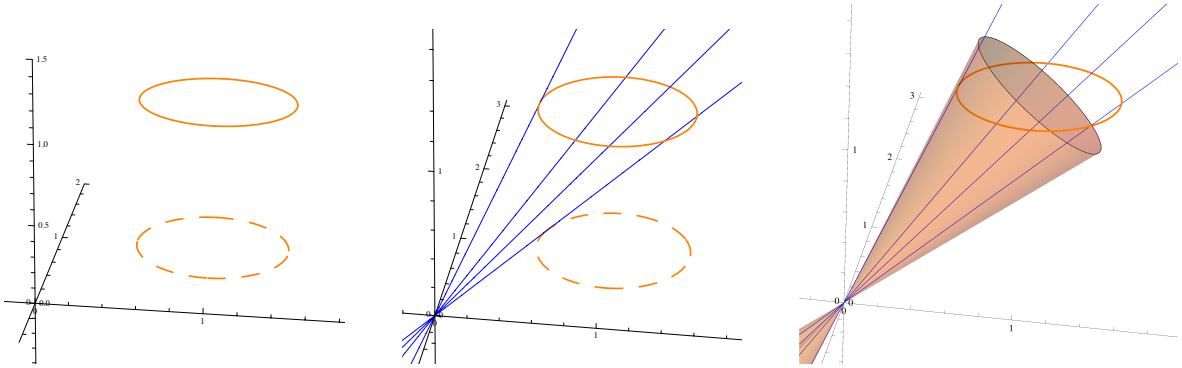


Figure 3: Let $K = 2$. The horizontal plane shows values of (z_1, z_2), while the vertical axis shows 'z_0'. The first plot shows the open ball in supp(Z) as a dashed circle, which is projected up into the plane $z_0 \equiv 1$, as a solid circle. We know all projections onto lines $\ell(1, z)$ in this 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 5, this implies that F and $\tilde{F}$ have the same moments. Thus $F = \tilde{F}$.

□

Lemma 4. 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. (See Landsberg (2012, page 115) who provides a statement, but no proof, of this result.)

Proof of lemma 4. Let $m(x)$ denote the Lebesgue measure on $\mathbb{R}^{\dim(x)}$. For a fixed $(x_1, \ldots, x_{K-1}) \in \mathbb{R}^{K-1}$, define

$$S_{x_1,\ldots,x_{K-1}} = \{x_K \in \mathbb{R} : p(x_1, \ldots, x_{K-1}, x_K) = 0\}.$$

Then

$$\begin{aligned} m(S) &= \int_{\mathbb{R}^K} \mathbb{1}_S(x)\, dm(x) \\ &= \int_{\mathbb{R}^{K-1}} \left[\int_{\mathbb{R}} \mathbb{1}_{S_{x_1,\ldots,x_{K-1}}}(x_K)\, dm(x_K) \right] dm(x_1, \ldots, x_{K-1}) \\ &= \int_{\mathbb{R}^{K-1}} \left[m(S_{x_1,\ldots,x_{K-1}}) \right] dm(x_1, \ldots, x_{K-1}) \\ &= 0. \end{aligned}$$

The second line follows by Fubini's theorem. The fourth line holds as follows: For a fixed $(x_1, \ldots, x_{K-1})$, the fundamental theorem of algebra implies that $p(x_1, \ldots, x_{K-1}, x_K)$ has finitely many roots x_K. Thus $S_{x_1,\ldots,x_{K-1}}$ is finite, and hence has measure zero. This holds for all $(x_1, \ldots, x_{K-1}) \in \mathbb{R}^{K-1}$, and hence $m(S_{x_1,\ldots,x_{K-1}})$, viewed as a function of $(x_1, \ldots, x_{K-1})$, is identically zero, and hence has zero integral. □

Lemma 5. 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 5. 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. □

Proof of proposition 2. I prove the result for π_1; the proof for π_2 is symmetric. I suppress conditioning on X everywhere.

1. First I show that A6 implies A5.1, all moments of π_1 are finite. For an arbitrary random K-vector Y with cdf F_Y, let

$$m_{j_1,\ldots,j_K} = \int |y_1|^{j_1} \cdots |y_K|^{j_K} dF_Y(y) \qquad j_1, \ldots, j_K \in \mathbb{N}$$

denote the absolute moments of Y. A generalized version of Hölder's inequality states that

$$m_{j_1,\ldots,j_K} \leq \prod_{k=1}^{K} m^{1/K}_{0\ldots,j_k \cdot K,\ldots,0} \qquad j_1, \ldots, j_K \in \mathbb{N},$$

where $m_{0\ldots,j_k \cdot K,\ldots,0} = \int |y_k|^{j_k \cdot K} dF_{Y_k}(y_k)$. (See Dunford and Schwartz 1958, page 527, exercise 2.) Thus, if all absolute moments of the coordinate random variables Y_j are finite, then all absolute moments of Y are finite.

Recall

$$\pi_1 \equiv (\pi_{11}, \pi_{12}, \pi_{13}) \equiv \left(\frac{U_1 + \gamma_1 U_2 + (\delta_1 + \gamma_1 \delta_2)' x}{1 - \gamma_1 \gamma_2}, \frac{\beta_1}{1 - \gamma_1 \gamma_2}, \frac{\gamma_1 \beta_2}{1 - \gamma_1 \gamma_2} \right).$$

A6.1 implies that

$$P\left(\frac{1}{|1 - \gamma_1 \gamma_2|} \leq \frac{1}{\tau} \right) = 1,$$

and hence

$$P\left(-\frac{1}{\tau} \leq \frac{1}{1 - \gamma_1 \gamma_2} \leq \frac{1}{\tau} \right) = 1.$$

For $n \in \mathbb{N}$, we have

$$\begin{aligned} \int |\pi_{12}|^n \, dF_{\pi_{12}} &= \int \left| \frac{\beta_1}{1 - \gamma_1 \gamma_2} \right|^n \, dF_{\beta_1, \gamma_1, \gamma_2} \\ &\leq \left| \frac{1}{\tau} \right|^n \int |\beta_1|^n \, dF_{\beta_1} \\ &< \infty, \end{aligned}$$

where the second line follows by A6.1 and the last line since β_1 has finite absolute moments by A6.3.

A6.2 implies that there is an M such that $\text{supp}(\gamma_i) \subseteq [-M, M]$ for $i = 1, 2$. This plus A6.1 show that

$$P\left(\frac{-M}{\tau} \leq \frac{\gamma_1}{1 - \gamma_1 \gamma_2} \leq \frac{M}{\tau} \right) = 1.$$

Hence

$$\begin{aligned}\int |\pi_{13}|^n \, dF_{\pi_{13}} &= \int \left| \frac{\gamma_1}{1-\gamma_1\gamma_2}\beta_2 \right|^n \, dF_{\beta_2,\gamma_1,\gamma_2} \\ &\leq \left| \frac{M}{\tau} \right|^n \int |\beta_2|^n \, dF_{\beta_2} \\ &< \infty.\end{aligned}$$

Next,

$$\begin{aligned}\int |\pi_{11}|^n \, dF_{\pi_{11}} &= \int \left| \frac{U_1 + \gamma_1 U_2 + (\delta_1 + \gamma_1\delta_2)'x}{1-\gamma_1\gamma_2} \right|^n dF_{\gamma_1,\gamma_2,U_1,U_2,\delta_1,\delta_2} \\ &\leq \left| \frac{1}{\tau} \right|^n \int |U_1 + \gamma_1 U_2 + \delta_1'x + \gamma_1\delta_2'x|^n dF_{\gamma_1,\gamma_2,U_1,U_2,\delta_1,\delta_2} \\ &\leq \left| \frac{1}{\tau} \right|^n \left(|U_1| + |\gamma_1 U_2| + |\delta_1'x| + |\gamma_1\delta_2'x| \right)^n dF_{\gamma_1,\gamma_2,U_1,U_2,\delta_1,\delta_2} \\ &\leq \left| \frac{1}{\tau} \right|^n \int \left(|U_1| + M|U_2| + \sum_{k=1}^K |\delta_{1k}| \cdot |x_k| + M \sum_{k=1}^K |\delta_{2k}| \cdot |x_k| \right)^n dF_{U_1,U_2,\delta_1,\delta_2}.\end{aligned}$$

Line 2 follows by A6.1. Line 3 follows by the triangle inequality. Line 4 follows by A6.2 and the triangle inequality again. This latter expression is only a function of absolute moments of $(U_1, U_2, \delta_1, \delta_2)$, which are all finite by A6.4. Thus the absolute moments of π_{11} are finite.

2. Next I show that A6 implies A5.2, π_1 is uniquely determined by its 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 π_{11}, π_{12}, and π_{13} are each separately uniquely determined by their moments.

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

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

   where the fourth line follows by A6.1 and the last line since the MGF of β_1 exists by A6.3. An analogous argument holds for $t < 0$. Thus the moment generating function of π_{12} exists and hence π_{12} is uniquely determined by its moments. An analogous argument shows that the moment generating function of π_{13} exists, using A6.2.

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

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

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 $U_1, U_2, \delta_1' x, \delta_2' x$, and then uses A6.1 and A6.2 to eliminate the γ_1 and γ_2's. That leaves us with a sum of the moment generating function of $(U_1, U_2, \gamma_1, \gamma_2)$ evaluated at various points. Each of these MGFs exists by assumption A6.4. Thus the moment generating function of π_{11} exists and hence π_{11} is uniquely determined by its moments.

□

Proposition 3. Suppose one of the following holds.

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

Then A6.1 and A1 hold.

Proof of proposition 3. Suppress conditioning on X. In all cases I will show that there is a $\tau \in (0, 1)$ such that $P[\gamma_1\gamma_2 \in (1-\tau, 1+\tau)] = 0$, which is equivalent to A6.1.

1. Since the sign of γ_1 and γ_2 are not equal with probability one, $P(\gamma_1\gamma_2 < 0) = 1$. Let τ be any number in $(0, 1)$. Then $1-\tau > 0$ and so $P(\gamma_1\gamma_2 \leq 1-\tau) = 1$. Hence $P[\gamma_1\gamma_2 \in (1-\tau, 1+\tau)] \leq P[\gamma_1\gamma_2 > 1-\tau] = 0$. Thus A6.1 holds.

2. By assumption there are $\tau_1, \tau_2 \in (0, 1)$ such that $P(|\gamma_1| \leq \tau_1) = 1$ and $P(|\gamma_2| \leq \tau_2) = 1$. 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 4.

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

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

   Hence $P[\gamma_1\gamma_2 \in (1-\tau, 1+\tau)] \leq P[\gamma_1\gamma_2 > 1-\tau] = 0$. Thus A6.1 holds.

3. Analogous to the previous case.

□

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

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

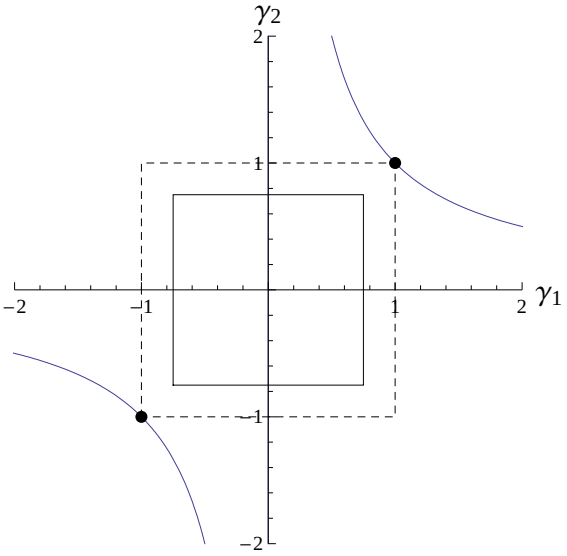


Figure 4: The solid rectangle is the boundary of $[-\tilde{\tau}, \tilde{\tau}]^2$. The dotted rectangle is the boundary of $[-1, 1]^2$. The line $\gamma_1\gamma_2 = 1$ is plotted.

and

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

□

Proof of theorem 3. The first part, $\inf_{\alpha \in \mathscr{A}_I} \|\hat{\alpha}_N - \alpha\|_{\mathscr{A}} = o_p(1)$, follows by verifying the conditions of theorem 3.1 in Chen et al. (2011). The weighted Hölder ball is $\|\cdot\|_c$-compact, which follows by modifying the proof of lemma A4 in Gallant and Nychka (1987) to use the Arzelá-Ascoli theorem instead of the Rellich-Kondrachov theorem, and then by applying lemma A.1 of Santos (2012). Since $\mathscr{F}$ is a closed subset of a compact set, it too is compact. Since the overall parameter space is compact, the penalty function assumptions 3.1.3 hold trivially. The sieve space restrictions 3.1.2 are assumed in E3 and the uniform convergence assumptions 3.1.4 are assumed in E4. Assumption 3.1.1(ii), upper semicontinuity, is implied by the full continuity assumption E2.2. The identified set is the inverse image of the closed set $\{E[\log p(Y \mid Z, X; \alpha_0)]\} \subseteq \mathbb{R}$ (i.e., a singleton set consisting of the true objective function value), under a continuous map, and hence $\mathscr{A}_I$ is closed in $\mathscr{A}$. Since $\mathscr{A}$ is compact, $\mathscr{A}_I$ is a closed subset of a compact set and hence is compact. Thus their assumption 3.1.1(iii) holds.

Next I show $\|\hat{f}_{\gamma_1|X} - f_{\gamma_1|X}\|_\infty = o_p(1)$. The proof is analogous for $\hat{f}_{\gamma_2|X}$. By the definition of $\|\cdot\|_{\mathscr{A}}$, the first part of this proof implies that

$$\inf_{f_{U,\Gamma|X} \in \mathscr{F}_I} \|\hat{f}_{U,\Gamma|X} - f_{U,\Gamma|X}\|_c = o_p(1) \qquad \text{and} \qquad \inf_{(b,d) \in \mathcal{B}_I \times \mathcal{D}_I} \|\widehat{(b,d)} - (b,d)\|_e = o_p(1),$$

where $\mathscr{A}_I = \mathcal{B}_I \times \mathcal{D}_I \times \mathscr{F}_I$ is the identified set. For the distribution of $\gamma_1 \mid X$, we have

$$\begin{aligned}
\|\hat{f}_{\gamma_1|X} - f_{\gamma_1|X}\|_\infty &= \sup_{(\gamma_1,x)} |\hat{f}_{\gamma_1|X}(\gamma_1 \mid x) - f_{\gamma_1|X}(\gamma_1 \mid x)| \\
&= \sup_{(\gamma_1,x)} \left| \int [\hat{f}_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x) - f_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x)]\, du_1\, du_2\, d\gamma_2 \right| \\
&\leq \sup_{(\gamma_1,x)} \int |\hat{f}_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x) - f_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x)|\, du_1\, du_2\, d\gamma_2 \\
&= \sup_{(\gamma_1,x)} \int |\hat{f}_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x) - f_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x)| \\
&\qquad \cdot \omega_c(u_1,u_2,\gamma_1,\gamma_2,x)\omega_c(u_1,u_2,\gamma_1,\gamma_2,x)^{-1}\, du_1\, du_2\, d\gamma_2 \\
&\leq \sup_{(\gamma_1,x)} \int \sup_{u_1,u_2,\gamma_1,\gamma_2,x} \{|\hat{f}_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x) - f_{U,\Gamma|X}(u_1,u_2,\gamma_1,\gamma_2 \mid x)| \\
&\qquad \cdot \omega_c(u_1,u_2,\gamma_1,\gamma_2,x)\}\omega_c(u_1,u_2,\gamma_1,\gamma_2,x)^{-1}\, du_1\, du_2\, d\gamma_2 \\
&= \|\hat{f}_{U,\Gamma|X} - f_{U,\Gamma|X}\|_c \sup_{(\gamma_1,x)} \int \omega_c(u_1,u_2,\gamma_1,\gamma_2,x)^{-1}\, du_1\, du_2\, d\gamma_2 \\
&\leq \|\hat{f}_{U,\Gamma|X} - f_{U,\Gamma|X}\|_c \int (1+u_1^2+u_2^2+\gamma_2^2)^{-\zeta_c}\, du_1\, du_2\, d\gamma_2 \\
&\leq \|\hat{f}_{U,\Gamma|X} - f_{U,\Gamma|X}\|_c \cdot C,
\end{aligned}$$

where $C < \infty$ since $\zeta_c > (4+K)/2$. Taking the infimum of $f_{U,\Gamma|X}$ over $\mathscr{F}_I$ of both sides gives

$$\|\hat{f}_{\gamma_1|X} - f_{\gamma_1|X}\|_\infty \leq \inf_{f_{U,\Gamma|X} \in \mathscr{F}_I} \|\hat{f}_{U,\Gamma|X} - f_{U,\Gamma|X}\|_c \cdot C$$

since $f_{\gamma_1|X}$ is identified. □