The Triangular Model with Random Coefficients

Stefan Hoderlein [*] Boston College
Hajo Holzmann[†] Marburg
Alexander Meister[‡] Rostock

June 9, 2015

The triangular model is a very popular way to capture endogeneity. In this model, an outcome is determined by an endogenous regressor, which in turn is caused by an instrument in a first stage. In this paper, we study the triangular model with random coefficients and exogenous regressors in both equations. We establish a profound non-identification result: the joint distribution of the random coefficients is not identified, implying that counterfactual outcomes are also not identified in general. This result continues to hold, if we confine ourselves to the joint distribution of coefficients in the outcome equation, except the one on the endogenous regressor. Identification continues to fail, even if we focus on means of random coefficients (implying that IV is generally biased), or let the instrument enter the first stage in a monotonic fashion. Based on this insight, we suggest an additional restriction that allows to point identify the distribution of random coefficients in the outcome equation. We extend this framework to cover the case where the regressors and instruments have limited support, and analyze semi- and nonparametric sample counterpart estimators in finite and large samples. Finally, we give an application of the framework to consumer demand.

Keywords: Random Coefficients, Endogeneity, Nonparametric, Identification, Radon Transform, Demand.

1. Introduction

The difference between causal effects and mere correlations is of crucial importance in microeconomics and is at the heart of the endogeneity issue. For instance, in consumer demand this type of difference arises naturally if unobservables like preferences over goods consumed today are correlated with factors like risk aversion that drive the level of overall total expenditure today. Heterogeneity is another common feature of microeconometric applications, meaning that causal effects vary widely across individuals. Staying in the consumer demand example, a small price change may result in a significant change in the behavior of some individuals while others leave their behavior largely unchanged. For many policy relevant questions, it is precisely this difference that is of interest. How causal effects in a heterogeneous population differ from a model that neither takes heterogeneity nor endogeneity into account is therefore a question of great importance.

[*]Department of Economics, Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467, USA, Tel. +1-617-552-6042. email: stefan_hoderlein@yahoo.com
[†]Department of Mathematics and Computer Science, Marburg University, Hans-Meerweinstr., 35032 Marburg, Germany, Tel. +49 -6421-2825454. email: holzmann@mathematik.uni-marburg.de
[‡]Institute for Mathematics, University of Rostock, 18051 Rostock, Germany, Tel. + 49 - 381-4986620. email: alexander.meister@uni-rostock.de

A very convenient tool to analyze this type of question is a linear correlated random coefficients model, as it embodies the notions of complex heterogeneity and endogeneity in a succinct, theory consistent way. In this model, the observable determinants of a scalar continuous outcome Y are related to this outcome by a structure that is linear in random coefficients B. Across the population, some of these determinants are now correlated with the random coefficients, while others are not. We denote the correlated (endogenous), resp., uncorrelated (exogenous), covariates by X, resp. W, For simplicity, we assume the former to be scalar. Since we are motivated by consumer demand applications, we will assume that X and W are continuously distributed; to fix ideas, think of total expenditure and prices.

The class of correlated random coefficient models (CRCs) we consider is then given by:

$$Y = B_0 + B_1 X + B_2' W,$$

where $B = (B_0, B_1, B_2')'$ is the vector of random coefficients. There are now basically two ways to deal with the endogeneity in the random coefficients. The first is by use of excluded exogenous variables Z that do not affect the outcome or the random coefficients directly, but which are correlated with X. The second is by use of panel data, or repeated cross sections. Examples for the first solution include Wooldridge (1997), Heckman and Vytlacil (1998), Florens et al (2008), Hoderlein, Klemelä, Mammen (2011), Masten (2013), and Masten and Torgovitsky (2014). All of these approaches employ instruments Z, and explicitly model the relationship between X and Z. The second route has been explored by, among many others, Chamberlain (1982, 1992), Arellano and Bonhomme (2013), and d'Haultfoeuille, Hoderlein and Sasaki (2013). Our approach falls into the former group, which itself is a subcategory of the greater category of triangular models, where the outcome depends on endogenous regressors which then in turn depend on variables Z that are excluded from the outcome equation, see, e.g., Imbens and Newey (2009) or Chesher (2003).

What distinguishes our paper from any of the previous contributions, with the notable exception of Masten (2013), is that we allow for several sources of unobserved heterogeneity in the relation between X and Z, and we do neither assume monotonicity of the first stage in a scalar heterogeneous factor, nor monotonicity in an instrumental variable Z. In fact, we specify the relationship between X and a vector $(Z', W')'$, henceforth called the first stage, fully coherently with the outcome equation as random coefficient model as well, i.e., the model is

$$\begin{aligned} Y &= B_0 + B_1 X + B_2' W, \\ X &= A_0 + A_1' Z + A_2' W, \end{aligned} \tag{1}$$

where $Z, A_1 \in \mathbb{R}^L$, $W, A_2, B_2 \in \mathbb{R}^S$, while the other quantities are scalar random variables. The variables Y, X, Z, W are observed, $A = (A_0, A_1', A_2')'$, $B = (B_0, B_1, B_2')'$ are unobserved random coefficients. As Kasy (2011) pointed out, in such a setup the random coefficients specification cannot simply be reduced to a scalar reduced form ("control function") heterogeneity factor in the first stage equation. As a consequence, we have to deal explicitly with this specification. We focus on high dimensional unobserved heterogeneity, since we believe it to be the most important feature of reality in many applications, and of linearity in random parameters as a reasonable first-order approximation on individual level. Compare this with the classical control function literature that allows for a nonlinear relation between Y and the regressors, at the expense of being only able to include a scalar unobserved factor only. Moreover, we include exogenous covariates W that appear in the first stage and the outcome equation - again fully consistently - through added terms $B_2' W$ and $A_2' W$ as well.

We shall always impose the following two basic assumptions. First, we assume that the random vector $(A', B')'$ has a continuous Lebesgue density f_{AB}. This continuity assumption will be maintained throughout the paper. While some of the theory could be extended to cover mass points with positive probability

(i.e., "types"), in line with the literature on random coefficient models (e.g., Hoderlein et al. (2011), Gautier and Kitamura (2013), Masten (2014)), we confine ourselves to this setup. Second, as key identifying restriction we will assume full independence of instruments and exogenous covariates:

Assumption 1 (*Independence*). $(Z',W')'$ and $(A',B')'$ are independent.

This assumption presents a natural strengthening of the common moment conditions found in the fixed coefficients linear model. This strengthening is necessary, because we allow for several sources of unobserved heterogeneity, and is again in line with the literature, in particular, any of the above references. In as far as we show nonidentification, the results would of course continue to hold under weaker form of independence.

Main Contributions. When studying this model in detail, we first uncover profound limitations in our ability to identify the object of interest, the density of the random coefficients, from the joint distribution of the observables. Consider the special case where X,Z and Y are scalar, and W is dropped from the model. Then we show by counterexample that the joint distribution of (A,B) is not identified, even if we focus on the subclass of smooth densities of compact support. It also continues to hold, if we consider the case where Z exerts a monotonic influence on X across the population, i.e., $A_1 > 0$ almost surely. Intuitively, the counterexample arises because it is impossible to map the three dimensional distribution of observables into the four dimensional distribution of parameters. More precisely, we show that there is a one-to-one mapping between the conditional characteristic function (ccf) of the data, and the characteristic function of the random coefficients on a three dimensional manifold only, and that there are several four dimensional densities that are compatible with this aspect of the characteristic function. Moreover, we can trace back the source of non-identification to the distribution of B_0; indeed, not even the mean of B_0 is point identified. Borrowing from the counterfactual notation of the treatment effects literature, this means that we cannot identify the distribution of $Y_x = B_0 + B_1 x$, for any x, in the absence of further assumptions This implies that we cannot identify analogs of the quantile treatment effect, i.e., $q_{Y_x}(\alpha) - q_{Y_{x-1}}(\alpha)$, where $q_S(\tau)$ is the τ-th quantile of a random variable S.

In the extended model including covariates W, the non-identification extends to the distribution and indeed the mean of B_2. While the marginals of B_0 and B_2 are of interest in their own right, note also that joint distributions are important. One may for instance be interested in the covariances between random parameters, say, between price and income effect. Moreover, they are of interest because some important economic quantities may be functionals of the joint distribution, e.g., the distribution of welfare effects in consumer demand.

These striking results suggest that we need to impose additional assumptions to identify the joint distribution of random coefficients, and even the marginals of f_{B_0} and f_{B_2}. We propose and discuss what we consider to be a natural assumption, namely that at least one random coefficient in the first stage equation is independent of the random coefficients in the outcome equation, an assumption that we justify in a consumer demand setup. Under this assumption, which actually includes the case where there is one fixed coefficient, we obtain a constructive point identification result that allows to represent the density of random coefficients in the outcome equation as an explicit functional of the distribution of the data, which may be used to construct a nonparametric sample counterparts estimator similar to that in Hoderlein et al. (2011), see the supplementary material.

However, the focus of the estimation part is to devise an estimator that incorporates the lessons learned from both the non-identification result, as well as the constructive identification result, in a parametric setup that is more relevant for applications. As was already mentioned, this paper is at least in parts motivated by applications in consumer demand. In this setup, and indeed many others, endogenous regressors

like prices and income (and instruments like tax or wage rates) can be thought of as approximately continuous, but they only vary on a bounded support. We consider thus in particular this latter issue. We show that the model is not identified by the means introduced before, and argue that this case requires the use of extrapolation strategies. We propose two such strategies, namely a parametric functional form and analiticity of the density of f_B. Since it is of particular relevance for applications, we focus in particular on the former, and we show how to construct an semi-parametric estimator that embodies the constructive nonparametric identification results, while at the same time being feasible in relatively high dimensional settings that arise frequently in applications. We also investigate the behavior of this estimators in large samples, and show that it achieves a parametric rate of convergence. Further, we analyze the behaviour of the linear instrumental variables estimator for the means of the random coefficients in the outcome equation. Finally, an application and a Monte Carlo study illustrate the performance of the proposed methodology.

Literature. Our model is closely related to index models with random coefficients. In particular, as already discussed, it is related to the work on the linear model in Beran and Hall (1992), Beran, Hall and Feuerverger (1996), Hoderlein et al. (2011), and Gautier and Hoderlein (2013). It also falls into the wider class of models analyzed in Fox and Gandhi (2009) and Lewbel and Pendakur (2013), who both analyze nonlinear random coefficient models, but the latter does not allow for endogeneity. The identification part is related to the innovative work in Masten (2013), who analyzes a fully simultaneous linear random coefficient system, which nests our model. There are several differences, though. Masten (2013) focuses on identification of the marginal distribution of B_1, and he does not establish non-identification of the entire distribution of B or A, nor of subsets thereof. Further, Masten (2013) does not provide conditions under which the rest of the model is identified.

Matzkin (2012) discusses the identification of the marginal distribution in a simultaneous equation model under additional constraints that make the model non-nested from ours. Chesher and Rosen (2013) discuss nonparametric identification in a general class of IV models that nests ours and achieve partial identification. Our approach in contrast adds structure and achieves point identification.

Since we have an explicit form for the first stage, it is instructive to compare it to triangular models, where Y is a function of X, and X is a function of Z. Most of the time, the outcome equation is left more general than a random coefficient models, at the expense of identifying (only) the average structural function, see Imbens and Newey (2009), and Kasy (2013), or some local average structural derivatives, see Hoderlein and Mammen (2009), also called local average response by Chamberlain (1982). The only random coefficients approaches we are aware of is the independent work of Masten and Torgovitsky (2013), who focus at the average random coefficient in a linear correlated random coefficient model with continuous outcome, and Hoderlein and Sherman (2011), who consider the same model with binary outcomes. All of these approaches, except the one proposed in Kasy (2013) employ control function residuals, and hence at least implicitly restrict the first stage heterogeneity to come from a scalar unobservable.

Finally, our motivation is partly driven by consumer demand, where heterogeneity plays an important role. Other than the large body of work reviewed above we would like to mention the recent work by Hausman and Newey (2013), Blundell, Kristensen and Matzkin (2013), see Matzkin (2007) and Lewbel (1999) for a review of earlier work.

Structure of the Paper. In Section 2.1, we provide a generic counterexample which shows that the joint distribution of A and B is not identified, and indeed we trace the lack of identification all the way back to EB_0. In Section 2.2 we show that the arguments extend to exogenous covariates. Further, we compute explicitely the linear IV parameter and show that it is biased for the mean of B. In Section 3.1, we establish constructive identification of the marginal distribution of B under fully-supported instruments

Z and exogenous regressors W in case of an additional independence assumption. An important extension relevant to applied work is discussed in Section 3.3: we consider the case of limited support of Z and W, and provide conditions that ensure identification. Without our additional assumption which allows for point identification, in Section 3.2 we provide bounds for the distribution function of B. The identification results lead to a semiparametric minimum-contrast estimator, the large sample properties of which are studied in Section 4. The finite sample properties of the estimator are studied through a Monte Carlo study in Section 5. Finally, in an extensive application to consumer demand, we first show how to identify the distribution of welfare effects, and then apply our estimator for both the random coefficients in the baseline model as well as the distribution of derived welfare effects to British consumption data, before an outlook concludes. Proofs are deferred to the appendix, while a supplement contains additional technical arguments.

2. Nonidentification of the distribution of the intercept B_0 and the slope B_2

Since the model in its general form (1) is quite involved, in order to study identification we proceed in several steps, illustrating issues in simpler models to keep them more transparent and then introducing extensions as they arise. In particular, we frequently use the simplest version of the model, which is given by

$$\begin{aligned} Y &= B_0 + B_1 X, \\ X &= A_0 + A_1 Z, \end{aligned} \tag{2}$$

where Y, X, Z are observed random scalars, and $A = (A_0, A_1)'$, $B = (B_0, B_1)'$ are unobserved random coefficients. Assumption 1 have to be adopted to this model in an obvious way. When analyzing the triangular RC model (2), it will often be convenient to pass to the reduced form model by inserting the second equation into the first one. This leads to

$$\begin{aligned} Y &= C_0 + C_1 Z, \\ X &= A_0 + A_1 Z. \end{aligned} \tag{3}$$

where $C = (C_0, C_1)$, $C_0 = B_0 + B_1 A_0$ and $C_1 = B_1 A_1$. In the following we shall write (A, B) and (A, C) instead of $(A', B')'$ and $(A', C')'$. In order to study the link between the distribution of (A, B) and (A, C) we introduce the mapping $\tau(a_0, a_1, b_0, b_1) := \big(a_0, a_1, b_0 + b_1 a_0, b_1 a_1\big)$. Note that the restriction of τ to the domain $\{a_1 \neq 0\}$ represents an invertible mapping to this set[1]. Indeed, we have $\tau^{-1}(a_0, a_1, c_0, c_1) = \big(a_0, a_1, c_0 - c_1 a_0/a_1, c_1/a_1\big)$. It follows that (A, C) has a Lebesgue density $f_{A,C}$ as well and that

$$f_{A,C}(a,c) = f_{A,B}\big(\tau^{-1}(a,c)\big)/|a_1|, \text{ and } f_{A,B}(a,b) = f_{A,C}\big(\tau(a,b)\big) \cdot |a_1|, \tag{4}$$

with Jacobian determinants $1/a_1$, and a_1, respectively.

2.1. Nonidentification of distribution of the intercept B_0

To illustrate the main issue, we first consider the basic model (2) and then show that the argument extends to model (1). Our results do not rely on support conditions or pathological counterexamples; indeed, densities can be extremely well behaved (e.g., analytic), and the results still continue to hold. Moreover, they hold even if A_1 is confined to be positive almost surely, and hence establish that it is not the case that monotonicity in the instrument is a sufficient condition for identification in the triangular RC

[1] Recall that (A, B) has a Lebesgue density $f_{A,B}$ so that $\{A_1 = 0\}$ is a null set.

model. Our (non-)identification argument is nonparametric, meaning that it may be possible to achieve identification through a fully parametric model, however, such results would rely exclusively on the parametric assumptions imposed, and would fully break down in case of a misspecified model.

To understand the source of nonidentification, let us first return to the basic model (2) and recall arguments in Masten (2013), who establishes identification of the density of B_1 in a setup that nests model (2). In case of fully supported Z, from (3) one can identify the joint distribution of (C_1, A_1), and hence also the distribution of $B_1 = C_1/A_1$. For B_0 however, the argument fails since the distribution of B_0 cannot be recovered from that of (C_0, A_0), as $C_0 = B_0 + B_1 A_0$, and neither can the distribution of counterfactual outcomes $Y_x = B_0 + B_1 x$, for all x, be identified (an this extends to B_2).

In the following, we will show this formally by counterexample, involving the reduced form (3). Let $\psi_{A,C}$ denote the characteristic function of (A', C'). By Assumption 1 we can relate the identified conditional characteristic function of (X, Y) given $Z = z$ to $\psi_{A,C}$ via

$$\begin{aligned} \psi_{X,Y|Z}(t_1, t_2|z) &:= E\big(\exp(it_1 X + it_2 Y)|Z = z\big) \\ &= E\exp\big(it_1(A_0 + A_1 z) + it_2(C_0 + C_1 z)\big) = \psi_{A,C}(t_1, t_1 z, t_2, t_2 z), \end{aligned} \tag{5}$$

where $z \in \operatorname{supp} Z$. The following lemma shows that this is actually all the information on (A', C') contained in the distribution of (X, Y, Z).

Lemma 2.1. *Let $(A', C')'$ and $(\tilde{A}', \tilde{C}')'$ be random vectors, independent of the exogenous variable Z which has a fixed distribution. Let (Y, X, Z) and $(\tilde{X}, \tilde{Y}, Z)$ be corresponding observed random variables from the model (3). If the characteristic functions $\psi_{A,C}$ and $\psi_{\tilde{A},\tilde{C}}$ of $(A', C')'$ and $(\tilde{A}', \tilde{C}')'$ coincide on the set*

$$\mathscr{S} = \big\{(t_1, t_1 z, t_2, t_2 z), \quad t_1, t_2 \in \mathbb{R}, z \in \operatorname{supp} Z\big\} \subseteq \mathbb{R}^4,$$

then the joint distributions of the observed variables (X, Y, Z) and $(\tilde{X}, \tilde{Y}, Z)$ will be equal.

As is explained below, this lemma is the basic building block of the following theorem, which shows that - in the absence of additional assumptions - the information provided by (X, Y, Z) does not suffice to identify neither the mean of B_0 nor, as a consequence, f_{B_0}.

Theorem 1. *Consider the triangular model (2) under Assumption 1. Suppose that all infinitely differentiable densities with compact support are admitted as joint density of $(A_0, A_1, B_0, B_1)'$. Then, the mean of B_0 cannot be identified from the distribution of the observations (X, Y, Z), even if Z is allowed to have full support.*

Since it is crucial to understand the limits of identification in this model, we now give the heuristics of the main steps involved in this result, while we defer the full formal construction to the Appendix, Section A.1. The basic intuition is that something three dimensional, i.e., the joint density of Y, X, Z cannot be used to identify something four dimensional, i.e., the joint density of (A, C) in general. Specifically, the set $\mathscr{S}$ in Lemma 2.1 is lower dimensional (in our example, it has three dimensions), and corresponds exactly to the set on which there is a one-to-one mapping between the object defined on this set, in our case aspects of the joint characteristic function of (A, C) and f_{YXZ}.

More formally, start out by noticing that for the polynomial $Q(u_0, u_1, v_0, v_1) = u_0 v_1 - u_1 v_0$, we have that

$$\mathscr{S} \subset \{(u_0, u_1, v_0, v_1) \in \mathbb{R}^4 : \quad Q\big(-i(u_0, u_1, v_0, v_1)\big) = 0\}. \tag{6}$$

Here, i is the imaginary unit with $i^2 = -1$, which we insert for a reason which will become clear immediately. Note that since Q is of degree two, we have $Q\big(-i(u_0, u_1, v_0, v_1)\big) = -Q(u_0, u_1, v_0, v_1)$.

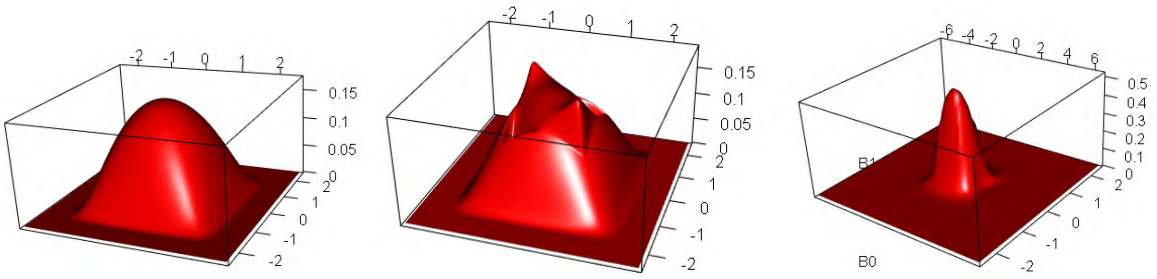

Figure 1: Bivariate marginal densities in the couterexample to identification. Left: Marginal of (A_0,A_1), Middle: Marginal of (A_0,C_1), Right: Marginal of (B_0,B_1)

Now, for a smooth function G_1 on $\mathbb{R}^4$ we let

$$r(a_0,a_1,c_0,c_1) = \big[\partial_{a_0}\partial_{c_1} - \partial_{a_1}\partial_{c_0} G_1\big](a_0,a_1,c_0,c_1) =: \big[Q\big(\partial_{a_0},\partial_{a_1},\partial_{c_0},\partial_{c_1}\big)G_1\big](a_0,a_1,c_0,c_1).$$

In Fourier space, differential operators turn into multiplication operators. More precisely, we have the formula

$$\begin{aligned}\big(\mathscr{F}_4 r\big)(u_0,u_1,v_0,v_1) =& \Big(\mathscr{F}_4\big[Q\big(\partial_{a_0},\partial_{a_1},\partial_{c_0},\partial_{c_1}\big)G_1\big]\Big)\\ =& Q\big(-i(u_0,u_1,v_0,v_1)\big)\,\big(\mathscr{F}_4 G_1\big)(u_0,u_1,v_0,v_1),\end{aligned}$$

where $\mathscr{F}_d$ denotes the d-dimensional Fourier transform.

By (6), this implies that the Fourier transform of the function r vanishes on the set $\mathscr{S}$, and since $0 \in \mathscr{S}$, in particular that $0 = \big(\mathscr{F}_4 r\big)(0) = \int r$. Therefore, we may add r to a given density G of (A,C), if G is chosen so that the resulting function remains non-negative, we obtain a density for (A,C) distinct from G for which, however, the Fourier transform coincide on $\mathscr{S}$ with that of G. By Lemma 2.1, based on the observed distribution of (X,Y,Z), one cannot discriminate between these densities. By change of variables, we can extend this negative result to (A,B), so that this joint distribution is not identified as well. Figure 1 illustrates this result. The left hand side corresponds to the two-dimensional marginals of G (which is a product density). In the center, the marginal of (A_0,C_1) (and (A_1,C_0)) under the joint density $G+r$ for (A,C) is plotted, the other two-dimensional marginals are equal to those under G itself. Finally, on the right we see the marginal of (B_0,B_1) when using $G+r$ for the distribution of (A,C), changing variables as in (4) and integrating out A.

To show that even the mean is not identified, we show that for these densities, even the means of B_0 differ. Since the difference of the densities is (a multiple of) r, this boils down to showing (see (4)) that

$$\int_{\mathbb{R}} b_0 \int_{\mathbb{R}^3} |a_1|\, r\big(a_0,a_1,b_0+a_1b_0,a_1b_1\big)\,da\,db_1\,db_0 \neq 0,$$

which may be accomplished using arguments from Fourier analysis, see Appendix, Section A.1.

2.2. Nonidentification of distribution of the slope B_2

The results on non-identification extend to the slope B_2 in model (1), where for simplicity, we still assume that there is one excluded instrument Z and add one exogenous covariate W. Parts of the intuition remain unchanged: It is still generically impossible to map a four dimensional distribution of observables into a six dimensional distribution of unobservables, as Theorem 8 in the Appendix, Section A.1 shows.

However, it is also the nature of the variation that is important, as the following special case without intercept shows:

Theorem 2. *Consider the triangular model (1) under Assumption 1, and suppose that $L = S = 1$. Then, neither the mean, nor the distribution of B_2 can be identified from the distribution of the observables (X,Y,Z,W), even if (Z,W) has full support, $A_0 = B_0 = 0$, and all infinitely differentiable densities of compact support are admitted as joint density of $(A_1,A_2,B_1,B_2)'$.*

The reason for this surprising finding is that the variation in W does not yield useful additional information. To see this, we use again the reduced form version of the general model,

$$\begin{aligned} Y &= C_0 + C_1 Z + C_2 W, \\ X &= A_0 + A_1 Z + A_2 W. \end{aligned}$$

where $C = (C_0, C_1, C_2)$, $C_0 = B_0 + B_1 A_0$, $C_2 = B_2 + B_1 A_2$ and $C_1 = B_1 A_1$. Note that C_2 has the same structural form as C_0, and our restriction translates to $A_0 = C_0 = 0$. Now, as in Lemma 2.1, all the information on the distribution of the random coefficients is contained in the conditional characteristic function of (X,Y) given (Z,W), and hence

$$\begin{aligned} E\Big(\exp\big(it_1 X + it_2 Y\big) \big| Z = z, W = w \Big) &= E\Big(\exp\big(i(t_1 z A_1 + t_1 w A_2 + t_2 z C_1 + t_2 w C_2)\big) \Big) \\ &= \psi_{A_1,A_2,C_1,C_2}(t_1 z, t_1 w, t_2 z, t_2 w), \end{aligned}$$

which identifies ψ_{A_1,A_2,C_1,C_2} over

$$\mathscr{S}' = \big\{ (t_1 z, t_1 w, t_2 z, t_2 w) \in \mathbb{R}^4 : \; t_1, t_2 \in \mathbb{R}, \; (z,w) \in \operatorname{supp}(Z,W) \big\},$$

and contains all the information from the joint distribution of the observations, in the sense analogous to Lemma 2.1. For the polynomial $Q(u_1,u_2,v_1,v_2) = u_1 v_2 - u_2 v_1$ as used in the proof of Theorem 1, we still have $\mathscr{S}' \subseteq \{(u_1,u_2,v_1,v_2)' \in \mathbb{R}^4 : \quad Q\big(-i(u_1,u_2,v_1,v_2)\big) = 0\}$, i.e., the variation in W is such that we cannot vary all four coordinates independently. Therefore, the same counterexample as in Theorem 1 applies.

2.3. Linear instrumental variables

Having established that it is impossible to point identify most parameters of interest in this model without further assumptions, a natural question that arises is what standard linear IV identifies and estimates in this model. In particular, we investigate the possible consistency of linear IV for the means EB_j of the random coefficients in the outcome equation. We consider the model of the previous section, with one excluded instrument Z and one additional covariate W. We assume the regressors to be centered, i.e., $EZ = EX = EW = 0$, otherwise the intercepts A_0 and B_0 have to be modified appropriately. The following result details what linear IV estimates:

Proposition 2.1. *Assume that (Y,X,Z,W) follow model (1) with Z and W being univariate, for which we maintain Assumption 1 (exogeneity of Z and W). If the random coefficients and the covariates Z,W have finite second moments, the covariates are centered, i.e. $EZ = EX = EW = 0$, and*

$$EA_1 \big(EZ^2 EW^2 - \big(EZW\big)^2 \big) \neq 0,$$

then linear IV estimates the population parameter

$$\mu_{IV} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & E[ZX] & E[ZW] \\ 0 & E[WX] & E[W^2] \end{pmatrix}^{-1} \begin{pmatrix} EY \\ E[YZ] \\ E[YW] \end{pmatrix} = \begin{pmatrix} EB_0 + E[A_0B_1] \\ E[A_1B_1]/EA_1 \\ EB_2 + E[A_2B_1] - E[A_1B_1]EA_2/EA_1 \end{pmatrix}$$

Apparently, the linear IV estimate is generally severely biased for all the means of B. The bias is not signed in general, and depends crucially on correlations of random coefficients across equations. As such, it is important to consider the structure in its entirety.

3. Identification of f_B

After these negative results, it is clear that generically additional identifying assumptions have to be introduced to achieve point identification, and we propose and discuss the marginal independence of A_1 from the coefficients B as a case in point. Note that this assumption still allows for A_0 and B to be arbitrarily dependent, as well as for A_1 and B to be dependent, conditional on A_0, but limits the direct dependence. We show how to achieve constructive point identification under this condition, first in the benign case where Z has full support (Section 3.1), and then in the case where Z has compact support (Section 3.3). To provide a concise and transparent exposition, we focus again on the basic model (2), and discuss the extension to the general model (1) less extensively.

3.1. Identification of f_B under full support

In the basic model (2), we first define the support condition formally:

Assumption 2. The exogenous variable Z in model (3) has full support $\mathbb{R}$.

To understand identification in this setup, we start out by recalling how the marginal density f_A of A is identified. Under Assumption 1, the identified conditional characteristic function of X given $Z = z$ relates to the characteristic function of the random coefficients A as follows[2]:

$$\mathscr{F}_1\big(f_{X|Z}\big)(t,z) = E\big(\exp\big(itX|Z=z\big)\big) = E\big(\exp\big(it(A_0+A_1z)\big)\big) = \big(\mathscr{F}_2 f_A\big)(t,tz).$$

Under the full support Assumption 2, and if $\mathscr{F}_2 f_A$ is integrable, we obtain by Fourier inversion that

$$f_A = T\Big(\mathscr{F}_1\big(f_{X|Z}\big)\Big), \tag{7}$$

where the operator T is defined by

$$\big(Tg\big)(a_0,a_1) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}}\int_{\mathbb{R}} |t| \exp\big(-it\,(a_0+a_1z)\big)\, g(t,z)\, dt\, dz, \tag{8}$$

which is well-defined for functions $g(t,z)$ which satisfy $\int_{\mathbb{R}}\int_{\mathbb{R}} |t|\,|g(t,z)|\,dtdz < \infty$, see the Appendix, Section A.2, for the proof.

After this preliminary step, we turn to the main identification question, i.e., identification of the marginal density f_B of B, which, as the example of Section 2 shows, will require additional structural assumptions. In particular, we will invoke the following assumption:

[2]As above, we let $\mathscr{F}_d$ denote the d-dimensional Fourier transform, and if $\mathscr{F}_1$ is applied to a function $g \in L_1(\mathbb{R}^2)$, it refers to Fourier transform w.r.t. the first coordinate when fixing the other coordinate.

Assumption 3 (*Independence and moment assumption*). Suppose that $B = (B_0, B_1)$ and A_1 are independent, and that A_1^{-1} is absolutely integrable.

This assumption obviously places structure on the dependence between the two random vectors A and B. We give examples of economic applications where this assumption is plausible. First, in our consumer demand application, B_1 reflects heterogeneity in individuals' reactions to changes in labor income, and ultimately, as argued above, the wage rate as exogenous driver of labor income under inelastic labor supply. Since this rate can be thought of as the costs of leisure, it reflects how heterogeneously individuals respond to these costs. In contrast, A_0 contains expected benefits of consumption (total expenditure). Since Y as total food expenditure comprises a large part of total expenditure, A_0 is likely correlated with the preferences that cause Y, in particular of course the random parameters B. This means that it is far less attractive to restrict the dependence between A_0 and B than between A_1 and B, as the relative taste for food and the reaction to labor supply cost factors are more likely to be independent.

Note, moreover, that Assumption 3 allows for A_0 and B to be arbitrarily dependent, as well as for A_1 and B to be dependent, conditional on A_0, and limits solely the direct dependence. In the example, this means that the heterogeneous reaction to the cost factors may well be correlated with the heterogeneous unobservable drivers of food budget choice, conditional on expectations about Y. Finally, we remark that this assumption is stronger than actually needed for our identification argument; we only need that $E\left[\left|A_1^{-1}\right| |B\right] = E\left[\left|A_1^{-1}\right|\right]$, as will be clear below. In sum, we feel that this assumption is defendable in many applications, however, this should not take away from the fact that this assumption amounts to placing structure on the unobservables - in the light of the non-identification results a necessary evil to achieve point identification.

To understand how this assumption allows us to relate the structural object f_B to the distribution of the data, more precisely the conditional characteristic function (ccf) of Y given X and Z, consider the following argument: Under Assumption 1, the ccf of Y given X and Z equals:

$$\begin{aligned} \mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z) &= E\left(\exp(itY)|X=x,Z=z\right) \\ &= E\left(\exp\left(it(B_0+B_1x)\right)|A_0+A_1z=x\right), \end{aligned} \tag{9}$$

where the second equality stems from the fact that $X = A_0 + A_1 Z$, which, after conditioning on $Z = z$, is solely a function of A, and hence independent of Z. It is easy to see that

$$\begin{aligned} &E\left(\exp\left(it(B_0+B_1x)\right)|A_0+A_1z=x\right) f_{X|Z}(x|z) \\ &\quad = \int_{\mathbb{R}^3} \exp\left(it(b_0+b_1x)\right) f_{B,A_0,A_1}(b, x-a_1z, a_1)\, da_1\, db_0\, db_1. \end{aligned}$$

Using the above two equations, applying the change of variables theorem, and integrating out z, we obtain:

$$\begin{aligned} &\int_{\mathbb{R}} \mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z)\, f_{X|Z}(x|z)\, dz \\ &\quad = \int_{\mathbb{R}^4} \exp\left(it(b_0+b_1x)\right) f_{B,A_0,A_1}(b, x-a_1z, a_1)\, dz\, da_1\, db_0\, db_1 \\ &\quad = \int_{\mathbb{R}^4} |a_1|^{-1} \exp\left(it(b_0+b_1x)\right) f_{B,A_0,A_1}(b, a_0, a_1)\, da_0\, da_1\, db_0\, db_1 \\ &\quad = E\left(\exp\left(it(B_0+B_1x)\right)|A_1|^{-1}\right). \end{aligned} \tag{10}$$

This is exactly where Assumption 3 comes into play: it allows to seperate out the factor $E\left[|A_1|^{-1}\right]$ from under the expectation, by making it not depend on B. As a minor remark, note also that Assumption 3

justifies the existence of the integral at the beginning of (10). Under the additional Assumption 3, we thus obtain that

$$\int_{\mathbb{R}} \mathscr{F}_1\big(f_{Y|X,Z}\big)(t|x,z)\, f_{X|Z}(x|z)\, dz = \big(\mathscr{F}_2 f_B\big)(t,tx)\, E|A_1|^{-1}. \tag{11}$$

As in (7), applying the operator T now yields the density f_B, and we get the following constructive point identification result:

Theorem 3. *In the triangular model (2), let Assumptions 1, 2 and 3, be true and assume that $\mathscr{F}_2 f_B$ is integrable.*

(i) Then the marginal density $f_B(b_0,b_1)$ of B is identified by

$$f_B(b_0,b_1) = T\Big(\int_{\mathbb{R}} \mathscr{F}_1\big(f_{Y|X,Z}\big)(t|x,z)\, f_{X|Z}(x|z)\, dz\Big)(b_0,b_1)\,\big(E|A_1|^{-1}\big)^{-1}, \tag{12}$$

where T, see (8), is applied w.r.t. the variables (t,x), and for every $x \in \mathbb{R}$, we have that

$$E|A_1|^{-1} = \int_{\mathbb{R}} f_{X|Z}(x|z)\, dz. \tag{13}$$

(ii) If, in addition, the smoothness Assumption 10 in Appendix A.2 is satisfied, we also have that

$$f_B(b_0,b_1) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}}\int_{\mathbb{R}} |t|\, \exp(-itb_0)\, \psi_{X,Y|Z}(-tb_1,t|z)\, dt\, dz\,\big(E|A_1|^{-1}\big)^{-1}. \tag{14}$$

Remark. The theorem shows that under additional assumptions, in particular Assumption 3, the joint density of B is identified because we can write it as an explicit functional of the distribution of the data. Note that, if all of A and B are independent there is no endogeneity. In this case,

$$\begin{aligned}\mathscr{F}_1\big(f_{Y|X,Z}\big)(t|x,z) =& E\big(\exp\big(it(B_0+B_1x)\big)|A_0+A_1z=x\big)\\ =& E\Big(\exp\big(it(B_0+B_1x)\big)\Big) = \mathscr{F}_1\big(f_{Y|X}\big)(t|x)\end{aligned}$$

does not depend on z. Thus, the conditional distribution of Y given $X=x, Z=z$ does not depend on $Z=z$ as well, and, as a consequence, Y is independent of $Z\,|\,X$. This provides a good basis for a test of exogeneity in our setup: If the estimated characteristic function depends on Z in a significant fashion (omission of variables test), we conclude that exogeneity is rejected. Furthermore, note that in the exogenous case where A and B are independent, $\mathscr{F}_1\big(f_{Y|X,Z}\big) = \mathscr{F}_1\big(f_{Y|X}\big)$ does not depend on Z. Therefore, observing (13) (see Appendix) and (11) reduces (12) to $f_B = T_1\big(\mathscr{F}_1\big(f_{Y|X}\big)\big)$, just as in (7). ◇

Remark. The second part (part (ii)) of the theorem shows that identification can also be achieved by considering the (identified) conditional characteristic function of (Y,X) given Z, which relates to the characteristic function of the reduced form coefficients as in (5). ◇

Now we turn to the extended model (1). The support assumptions need to be modified in an obvious way.

Assumption 4. In model (1), the exogenous vector $(Z',W')'$ has full support $\mathbb{R}^{L+S}$.

Next, in addition to the maintained assumption of instrument independence, we again need to place additional conditions on the dependence structure of the random coefficient vector. As it turns out, relatively speaking, these conditions are much less restrictive than in the simple model. We only need B to be independent of one of the slope coefficients, it can be arbitrarily correlated with all others as well as the intercept. To state this formally, for a vector $z=(z_1,\dots,z_L)' \in \mathbb{R}^L$ we write $z_{-1}=(z_2,\dots,z_L)'$, so that $Z=(Z_1,Z_{-1}')'$ and $A_1=(A_{1,1},A_{1,-1}')'$. The modified additional independence assumption is then:

Assumption 5 (*Independence and moment assumption*). Suppose that B and $A_{1,1}$ are independent, and that $A_{1,1}^{-1}$ is integrable.

The interpretation of this assumption is very similar to the above, with the caveat that $A_{1,-1}$ may be arbitrarily related to B. We remark that full independence is again stronger than necessary, and we in fact only require a conditional expectation to not depend on B, but because the weakening is not economically important, we desist from this greater generality. Moreover, it is automatically satisfied, if one coefficient is nonrandom.

To now state our main result, we define the operator T_K by

$$\big(T_K g\big)(s,x) = \frac{1}{(2\pi)^{K+1}} \int_{\mathbb{R}^{1+K}} |t|^K \exp\big(-it\,(s+x'v)\big)\, g(t,v)\, dt\, dv, \qquad s \in \mathbb{R},\ x \in \mathbb{R}^K,$$

where g satisfies $\int_{\mathbb{R}^{1+K}} |t|^K\, |g(t,v)|\, dt\, dv < \infty$. Our constructive identification result is then as follows:

Theorem 4. *Under Assumptions 1, 4 and 5 in the triangular model (1), if $\mathscr{F}_{2+S} f_B$ is integrable, the marginal density $f_B(b)$ of B is identified as*

$$f_B(b) = C \cdot T_{S+1}\Big(\int_{\mathbb{R}^L} \mathscr{F}_1\big(f_{Y|X,Z,W}\big)(t|x,z,w)\, f_{X|Z,W}(x|z,w)\, f_{Z_{-1}}(z_{-1})\, dz \Big)(b),$$

where T_{S+1} is applied w.r.t. the variables $\big(t, \big(x,w'\big)\big)'$, and where for every $w \in \mathbb{R}^S$, $z_{-1} \in \mathbb{R}^{L-1}$, $x \in \mathbb{R}$,

$$C^{-1} := E|A_{1,1}|^{-1} = \int_{\mathbb{R}} f_{X|Z,W}(x|z,w)\, dz_1. \tag{15}$$

This result may be seen as natural extension of the result in the special case without W, and a one dimensional Z. In particular, the change of variables step involves only one variable, and it is precisely this one variable whose coefficient appears then in the denominator. Otherwise, the basic structure is largely identical.

Remark. A natural conjecture is that assumption 3 is sufficient to make IV consistent. However, the limit only simplifies to

$$\mu_{IV,ind} = \begin{pmatrix} EB_0 + E\left[B_1 A_0\right] \\ EB_1 \\ EB_2 + \mathrm{Cov}\left(A_2, B_1\right) \end{pmatrix}$$

Thus, IV will become consistent for the mean of the coefficient of the endogenous regressor X, but will generally remain biased for the means of the coefficients of the exogenous regressor W and the intercept. In the application we deal with bivariate W, and in the supplementary material show that the above analysis extends to this situation. Importantly, however, if the coefficient on W in the first stage equation is nonrandom, IV is also consistent for EB_2. The inconsistency thus hinges on both equations having random coefficients.

3.2. Partial Identification

Going back to the simple model (2), if we drop Assumption 3, then our results in Section 2 show in particular nonidentification of the marginal distribution of B_0, while the marginal distribution of B_1 remains identified.

It is, however, possible to derive bounds on the distribution function of B_0, and hence also on the joint distribution function of (B_0, B_1). To this end, if we follow the argument starting in (9) leading to (10)

with e^{itY} replaced by $1_{Y\leq t}$, we obtain

$$\int_{\mathbb{R}} E\big(1_{Y\leq t}\,|X=x,Z=z\big)\,f_{X|Z}(x|z)\,dz = E\Big(1_{B_0+B_1x\leq t}\,|A_1|^{-1}\Big).$$

In case $0\in \operatorname{supp} X$, this yields the identification of

$$F(t) := \int_{\mathbb{R}} E\big(1_{Y\leq t}\,|X=0,Z=z\big)\,f_{X|Z}(0|z)\,dz = E\Big(1_{B_0\leq t}\,|A_1|^{-1}\Big). \tag{16}$$

The right hand side can be used to bound the distribution function $F_{B_0}(t)$ of B_0 in terms of $F(t)$ under additional assumptions.

Proposition 3.1. *In model (2) with $0\in\operatorname{supp} X$, define the identified function $F(t)$ as in (16).*
(i) If $0<c_{A_1}\leq A_1\leq C_{A_1}$ a.s., then $c_{A_1}F(t)\leq F_{B_0}(t)\leq C_{A_1}F(t)$ for all $t\in\mathbb{R}$.
(ii) If $E(|A_1|^{-p})<\infty$ for some $p>1$, then

$$\big(F(t)^p/E(|A_1|^{-p})\big)^{1/(p-1)}\leq F_{B_0}(t),\qquad t\in\mathbb{R}.$$

(iii) If $E(|A_1|^{p})<\infty$ for some $p>1$, then

$$F_{B_0}(t)\leq \big(F(t)\big)^{p/(p+1)}\,\big(E(|A_1|^{p})\big)^{1/(p+1)},\qquad t\in\mathbb{R}.$$

Note that (i) arises as the limit case for $p\to\infty$ from (ii) and (iii). The bounds in the proposition are based on the Hölder's inequality, applied to $|A_1|^{-1}$ and $1_{B_0\leq t}$. Equality only occurs in Hölder's inequality if (powers of) these random variables are linearly dependent, which cannot happen if A_1 has a Lebesgue density. In the limit case where $|A_1|^{-1}$ tends to a binary variable, linearly dependent with $1_{B_0\leq t}$, however, it may arise.

Given bounds $F_{l,B_0}(t)\leq F_{B_0}(t)\leq F_{u,B_0}(t)$, from the Fréchet-Hoeffding bounds we obtain bounds on the joint distribution function $F_{B_0,B_1}(t_0,t_1)$ as follows:

$$\begin{aligned}\max\big(F_{l,B_0}(t_0)+F_{B_1}(t_1)-1,0\big)&\leq \max\big(F_{B_0}(t_0)+F_{B_1}(t_1)-1,0\big)\\&\leq F_{B_0,B_1}(t_0,t_1)\\&\leq \min\big(F_{B_0}(t_0),F_{B_1}(t_1)\big)\leq \min\big(F_{u,B_0}(t_0),F_{B_1}(t_1)\big).\end{aligned}$$

3.3. Identification under limited support

In applications, it is not plausible that continuous regressors may vary over the whole real line. In this section, we hence extend our approach to deal with the situation where Z has compact support. The first important observation is that we can not - in general - follow the approach which led to Theorem 3. Indeed, in this case, the identifying relation (9) holds (only) for $z\in\operatorname{supp} Z$. If A_0 has infinite support, we cannot integrate out z in (10) over the whole of $\mathbb{R}$ in order to get rid of A_0. There are two possible routes that follow from this observation: the first is to limit the support of A in conjunction with assuming functional form structure on B; the second is to invoke assumptions that allow to extrapolate from $\operatorname{supp} Z$.[3] Both strategies have their merits and problems, and both strategies have precedents in the econometric literature. We will now discuss them in turn.

[3] As will become evident below, restricting the support of B will not suffice to in general to obtain point identification, even if the support of A is restricted. In order to extrapolate, one can either assume functional form structure, or one may assume analytic continuation structure. In the latter case, one can also obtain the result without support restrictions, which is why we follow this strategy in the second approach.

Support Restrictions. As it turns out, restricting the support of the random coefficients in the first stage equation allows to use arguments from the previous subsection. To see this, consider the extended model (1), but for simplicity we assume to have a univariate Z (so $L = 1$). However, we allow for the exogenous covariates W to be multivariate. Formally, our support restriction will be as follows:

Assumption 6. There exist pairs $(x, w') \in \operatorname{supp}(X, W)$ for which

$$\operatorname{supp}\Big(\frac{x - A_0 - A_2' w}{A_1}\Big) \subseteq \operatorname{supp}\big(Z|W = w\big) =: \mathscr{S}_{Z,w} \tag{17}$$

To understand this assumption, let Z have bounded support in the sense that $\operatorname{supp}\big(Z|W = w\big) = [z_l, z_u]$. Moreover, assume that w is such that $\operatorname{supp}(A_0 + A_2' w, A_1) \subset [a_l, a_u] \times [a_{1,l}, a_{1,u}]$, where $a_{1,l} > 0$, i.e., for any w the support of A is contained in the same rectangle. For an $x \in [a_u, a_u + a_{1,u} z_u]$, it then holds that

$$\operatorname{supp}\big((x - A_0 - A_2' w)/A_1\big) \subset \big[(x - a_u)/a_{1,u}, (x - a_l)/a_{1,l}\big].$$

To obtain $\operatorname{supp}\big((x - A_0 - A_2' w)/A_1\big) \subset \operatorname{supp}(Z|W = w)$ for such an x, we require that $z_l \le (x - a_u)/a_{1,u}$ and $(x - a_l)/a_{1,l} \le z_u$. Thus, for all $x \in \operatorname{supp}(X|W = w)$ with $a_{1,u} z_l + a_u \le x \le a_{1,l} z_u + a_l$, (17) is satisfied. Hence, since

$$\operatorname{supp}(X|W = w) \subset \big[a_l + \min\big(a_{1,l} z_l, a_{1,u} z_l\big), a_u + \max\big(a_{1,u} z_u, a_{1,l} z_u\big)\big],$$

if the support of $Z|W = w$ is sufficiently large as compared to that of $(A_0 + A_2' w, A_1)$, (17) will be satisfied for an interval of x values. As such, the limited variation in Z allows to still apprehend all values of A, which is the core effect of the support restriction.

Theorem 5. *Consider the triangular model (1) in case of a univariate Z. Impose the Assumptions 1 and 5. Then, for all $t \in \mathbb{R}$ and all $(x, w') \in \operatorname{supp}(X, W)$ which satisfy (17), the following holds*

$$\big(\mathscr{F} f_B\big)(t, tx, tw) = \big(E|A_1|^{-1}\big)^{-1} \int_{\mathscr{S}_{Z,w}} \mathscr{F}_1\big(f_{Y|X,Z,W}\big)(t|x, z, w)\, f_{X|Z,W}(x|z, w)\, dz. \tag{18}$$

Setting $t = 0$ yields in particular that

$$E|A_1|^{-1} = \int_{\mathscr{S}_{Z,w}} f_{X|Z,W}(x|z, w)\, dz \tag{19}$$

Discussion of Theorem 5. While identification of $\big(\mathscr{F} f_B\big)(t, tx, tw)$ for all t and for (x, w) varying in an open set does not suffice to identify f_B fully nonparametrically, it will typically suffice to identify a parametric model, such as the parameters of the normal distribution. Therefore, (18) can (and will be used below) to construct a minimum-distance parametric estimator.

Based on Theorem 5, one could also use analytic continuation arguments to identify f_B in a class of models with bounded support (and hence with analytic characteristic function). However, it turns out that the support restriction (17) is not necessary for this purpose, and we follow this second approach in greater generality in the next section. ◇

Analytic Continuation. While parts of the previous subsection already involved extrapolation arguments in a parametric form combined with support constraints on the density of random coefficients, we will now turn to a strategy that allows for quite general nonparametric identification of f_B, even with compactly supported Z and without the (potentially restrictive) support restriction (17) on the random coefficients, by using analytic continuation arguments.

We still need to assume that the random coefficients do not have heavy tails, as made precise in the following assumption.

Assumption 7. In model (3), all the absolute moments of A_1 and $C_1 = B_1 A_1$ are finite and satisfy

$$\lim_{k\to\infty} \frac{d^k}{k!}\left(E|A_1|^k + E|C_1|^k\right) = 0\,,$$

for all fixed $d \in (0,\infty)$.

This assumption is in particular satisfied if A_1 and B_1 have compact support.

Theorem 6. *We consider the triangular model (2) under the Assumptions 1, 3, and 7 and Assumption 10 (see Appendix A.2). If the support of Z contains an open interval, and if $\mathscr{F}_2 f_B$ is integrable, then the density f_B of (B_0, B_1) is identified.*

This identification result is more abstract than in Theorem 3, since no explicit inversion formula is obtained. Nevertheless, it can be used to construct a nonparametric estimator, and in the technical supplement to this paper, we outline and analyze such an estimator. Moreover, we remark that an identification result under limited support in model (1), based on analytic continuation, can be obtained similarly.

4. Semiparametric estimation

Next we discuss how the insights obtained from the identification results translate into estimation approaches and estimation theory.

Since simple parametric estimation, e.g. assuming that f_{AB} is the multivariate normal distribution, runs into the problem that the estimator relies on elements which are not nonparametrically identified, we do not recommend this route. Instead, we show how to use the identification results to construct root-n-consistent semiparametric estimators. We think of these types of estimators as being most relevant for applications. In addition, because of the greater relevance of the limited support case for applications, we concentrate on this case. In the supplementary material, we also discuss nonparametric estimation, both for full and limited support. While arguably less relevant in practice, we think of this topic as important as it illustrates how the insights from identification are reflected in the structure of a sample counterpart estimator, and how the various parts affect the behavior of the estimation problem.

In order to keep the technical arguments as transparent and simple as possible, we develop asymptotic theory only in the simple triangular model (2), but also show how the estimators may be extended to include exogenous covariates W. Throughout, we shall maintain the basic Assumptions 1 and 3.

Our semiparametric estimator will rely on the identification results in Theorem 5. We start the estimation of the scaling factor $E|A_1|^{-1}$. Let us specialize the compact support assumption used in this section as follows.

Assumption 8. Assume that Z has support $[-1,1]$, and has a density f_Z with $f_Z(z) \geq c_Z$ for all $z \in [-1,1]$ for some $c_Z > 0$.

Note that the interval $[-1,1]$ is chosen for convenience. Moreover, the lower bound on the density of f_Z is no major restriction; its existence may be justified by considering a subset of the support of Z.

We additionally require the support restriction (17) for an interval $I \subseteq \operatorname{supp} X$, i.e. that

$$\operatorname{supp}\left(\frac{x - A_0}{A_1}\right) \subseteq \operatorname{supp} Z = [-1,1] \quad \forall x \in I\,. \tag{20}$$

From Theorem 5, for all $x \in I$, $E|A_1|^{-1} = \int_{-1}^{1} f_{X|Z}(x|z)\,dz$. For the given interval I choose a bounded weight function $v : \mathbb{R} \to (0,\infty)$ with $\operatorname{supp} v \subseteq I$ and $\int_I v = 1$. Observe that

$$E|A_1|^{-1} = \int_I v(x) \int_{-1}^{1} f_{X|Z}(x|z)\,dz\,dx,$$

which motivates the following Priestley-Chao-type estimator

$$\hat{a}_{I,n} = \sum_{j=1}^{n-1} v(X_{(j)}) \left(Z_{(j+1)} - Z_{(j)}\right), \tag{21}$$

where we denote by $(X_{(j)}, Y_{(j)}, Z_{(j)})$, $j = 1,\dots,n$, the sample sorted according to $Z_{(1)} < \cdots < Z_{(n)}$.

For the following result, we need an additional regularity condition, in particular an assumption on the boundedness and smoothness of the density f_A. This Assumption 11 is deferred to the Appendix, Section B, for brevity of exposition; we will do likewise in the subsequent theorem.

Proposition 4.1. *Given Assumptions 8 and 11, for an interval $I \subseteq \operatorname{supp} X$ for which (20) is satisfied, we have that*

$$E\left(\hat{a}_{I;n} - E|A_1|^{-1}\right)^2 = \mathcal{O}\left(n^{-1}\right).$$

This result establishes that the squared distance between the estimator of the scaling constant and the true parameter decreases at the parametric rate.

Next, we construct an estimator for the marginal density of f_B in a parametric model $\left\{f_{B,\theta} : \theta \in \Theta\right\}$. Note that we do not assume a parametric model for all random coefficients, i.e., f_{AB}, but only for f_B - our model is thus semiparametric. There are several reasons for such a semiparametric approach in contrast to a fully parametric one: First, we are robust against possible misspecifications in the parametric form of the distribution of (A, B_0) as well as of (A_0, A_1). Second and more importantly, a fully parametric specification would rely on and hence require identification of the joint distribution of (A, B) (given the additional Assumption 3). Our identification results do not establish this, and in fact we conjecture that such extended identification is not valid.

To proceed, given the nonparametric estimator for the scaling constant $E|A_1|^{-1}$, we now want to estimate the density of random coefficients semi-parametrically in the case of bounded Z. To this end, suppose that the interval $I \subseteq \operatorname{supp} X$ satisfies (20), and further that Assumption 8 is satisfied. By (18),

$$\int_{-1}^{1} \int \exp(ity) \int_I f_{Y,X|Z}(y,x|z)dx\,dy\,dz = E|A_1|^{-1} \cdot \int_I (\mathscr{F} f_B)(t,tx)dx. \tag{22}$$

Note that this equation holds even in the absence of any functional form assumption. Suppose now that $f_B = f_{B,\theta_0}$ belongs to the parametric family of models $\left\{f_{B,\theta} : \theta \in \Theta\right\}$, where $\Theta \subset \mathbb{R}^d$ is a d-dimensional bounded cuboid. The procedure we propose is now as follows: We estimate the left hand side of (22) nonparametrically by

$$\hat{\Phi}_n(t,I) := \sum_{j=1}^{n-1} \exp(itY_{(j)}) 1_I(X_{(j)}) \cdot \left(Z_{(j+1)} - Z_{(j)}\right),$$

and compare it with the right hand side that features a parametric specification, $f_{B,\theta}$. This comparison defines an appropriate contrast (or distance) that we use to estimate θ_0. For $\theta \in \Theta$ and $t \in \mathbb{R}$ we let

$$\Phi(\theta,t,I) := \int_I (\mathscr{F} f_{B,\theta})(t,tx)dx.$$

To define the contrast, let ν be a probability measure on $\mathbb{R}$, and let $I_1,\dots,I_q$ be finitely many (distinct) intervals which satisfy (20). For bounded functions $\Phi_1(t,I_j)$ and $\Phi_2(t,I_j)$, $t\in\mathbb{R}$, $j=1,\dots,q$, we set

$$\|\Phi_1(\cdot)-\Phi_2(\cdot)\|^2_{\nu;q} := \frac{1}{q}\sum_{j=1}^{q}\int_{\mathbb{R}}\left|\Phi_1(t,I_j)-\Phi_2(t,I_j)\right|^2 d\nu(t), \tag{23}$$

and note that $\|\cdot\|_{\nu;q}$ defines a seminorm. Let $\hat{a}_{I;n}$ be a $\sqrt{n}$-consistent estimator for $E|A_1|^{-1}$, as given in Proposition 4.1. Taking into account (22), we choose our estimator as a minimizer in θ of the functional

$$\theta \mapsto \|\hat{\Phi}_n(\cdot)-\hat{a}_{I;n}\cdot\Phi(\theta,\cdot)\|^2_{\nu;q}. \tag{24}$$

In order to reduce technicalities, we use the method of discrete approximation for M-estimators. More specifically, for some constant $b>0$, define the grid

$$\Theta_n := \Theta\cap G_{b,n}, \qquad G_{b,n} := \left\{jbn^{-1/2} : j\in\mathbb{Z}\right\}^d,$$

and choose $\hat{\theta}_n$ as a (any) minimizer of (24) over the grid Θ_n. The grid and hence the estimator depend on the parameter b which adds additional flexibility, however, the result below is true for any fixed $b>0$, and we therefore suppress the b in the notation of the estimator $\hat{\theta}_n$ and the statement of the theorem.

Two main ingredients are required so that $\hat{\theta}_n$ achieves the parametric rate. First, the empirical version $\hat{\Phi}_n(t,I)$ converges in an appropriate sense to the asymptotic one $E|A_1|^{-1}\Phi(\theta_0,t,I)$ at the parametric rate. Second, the asymptotic contrast between distinct parameters needs to be of the same order as the Euclidean distance between those parameters, in the sense of the following assumption.

Assumption 9. There exist intervals $I_1,\dots,I_q$ satisfying (20) and a probability measure ν, such that

$$\left(E|A_1|^{-1}\right)^{-1} c_{\Theta,0}\left\|\theta-\theta'\right\|^2 \le \|\Phi(\theta,\cdot)-\Phi(\theta',\cdot)\|^2_{\nu;q} \le \left(E|A_1|^{-1}\right)^{-1} c_{\Theta,1}\left\|\theta-\theta'\right\|^2$$

for all $\theta,\theta'\in\Theta$ with some uniform constants $c_{\Theta,j}$, $j=0,1$.

Note that the scaling constant $\left(E|A_1|^{-1}\right)^{-1}$ could be included in the $c_{\Theta,j}$, $j=0,1$, but the above notation will be convenient in the proof of the theorem below.

Example (Bivariate normal distribution). While the above assumption is reasonable, showing it for specific parametric models may be quite involved. In the technical supplement Section E, we give a fully rigorous proof of the validity of Assumption 9 in the most important special case, i.e., a bivariate normal distribution, with two distinct values of t and three disjoint intervals I_j. In practice, other choices of the weighting measure are more convenient. If we choose $d\nu(t)$ to be centered Gaussian with standard deviation s, then after dropping terms not depending on the parameters as well as constant factors, we need to minimize

$$\begin{aligned} &M(\mu,\Sigma)\\ &=\sum_{p=1}^{q}\Big(\hat{a}_{I,n}\int_{I_p}\int_{I_p}\varphi\big(0;\mu_1(x_2-x_1),\big((1,x_1)\Sigma(1,x_1)'+(1,x_2)\Sigma(1,x_2)'^{-2}\big)^{-1}\big)\,dx_1\,dx_2\\ &-2\sum_{j=1}^{n-1}\int_{I_p}\varphi\big(0;Y_{(j)}-\mu_0-\mu_1 x,\big((1,x)\Sigma(1,x)'^{-2}\big)^{-1}\big)\,1_{I_p}(X_{(j)})\cdot\big(Z_{(j+1)}-Z_{(j)}\big)dx\Big). \end{aligned} \tag{25}$$

◇

Here, $\hat{a}_{I,n}$ be the estimator of the scale constant based on the full interval I, the I_p partition I, and $\varphi(t;\eta,\tau^2)$ the density of $N(\eta,\tau^2)$.

Under the above assumption, and again a regularity condition (Assumption 12) to be found in the appendix, we obtain:

Theorem 7. *Suppose that the marginal density $f_B = f_{B,\theta_0}$ belongs to the parametric model $\{f_{B,\theta} : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ is a d-dimensional bounded cuboid. Given Assumptions 1, 3, 8, 9 and 12, the estimator $\hat{\theta}_n$ satisfies*

$$\left\|\hat{\theta}_n - \theta_0\right\| = \mathcal{O}_P(n^{-1/2}).$$

This result establishes that our semiparametric estimator indeed achieves the parametric rate. As such, we may be cautiously optimistic that the estimator performs reasonably well in datasets of the size commonly found in applications.

Finally, we briefly discuss how to extend the estimator to model (1) which includes exogenous covariates W. We maintain the identification Assumption 5, and restrict ourselves to univariate W and Z.

Assume that the support $\mathcal{S}_{Z,w} = I_Z$ of Z given $W = w$ is a compact interval, independent of w (the conditional distribution itself may depend on w), and that the support of W is the compact interval I_W. Further, impose the support restriction (17) for a rectangle $I_X \times I_W \subseteq \operatorname{supp}(X,W)$, i.e.

$$\operatorname{supp}\left(\frac{x - A_0 - wA_2}{A_1}\right) \subseteq \operatorname{supp}\big(Z|W=w\big) = I_Z\,, \quad \forall\,(x,w)' \in I_X \times I_W\,. \tag{26}$$

Moreover, for the joint density of (Z,W) we assume that $f_{Z,W}(z,w) \geq c > 0$ for all $z \in [-1,1]$, $w \in I_W$, for some $c > 0$.

From Theorem 5, for all $(x,w)' \in I_X \times I_W$,

$$E|A_1|^{-1} = \int_{-1}^{1} f_{X|Z,W}(x|z,w)\,dz.$$

Choose a bounded weight function $v : \mathbb{R}^2 \to (0,\infty)$ with $\operatorname{supp} v \subseteq I_X \times I_W$ and $\int v = 1$, then we have that

$$E|A_1|^{-1} = \int_{I_X \times I_W} v(x,w) \int_{-1}^{1} f_{X|Z,W}(x|z,w)\,dz\,dx\,dw.$$

To generalize the weights $Z_{(j+1)} - Z_{(j)}$ from the situation without W, we recommand to use the following Priestly-Chao type weights

$$\lambda_{j,PC} = \operatorname{Area}\Big\{(z,w) \in I_Z \times I_W \,:\, \big|(z,w) - (Z_j,W_j)\big| \leq \big|(z,w) - (Z_k,W_k)\big|,\, \forall k = 1,\ldots,n\Big\},$$

$j = 1,\ldots,n$, where Area denotes the Lebesgue area. Actually, in the univariate situation without W, this corresponds to the weights $\lambda_{j,PC} = \big(Z_{(j+1)} - Z_{(j-1)}\big)/2$, which gives the same results asymptotically as $Z_{(j+1)} - Z_{(j)}$ as chosen previously. In the multivariate situation it is hard to compute the $\lambda_{j,PC}$ analytically. However, it is straightforward to approximate them using Monte Carlo: for given $N \in \mathbb{N}$ (we use $N = 200$ in the simulation section), generate i.i.d. $U_1,\ldots,U_{n\cdot N}$, uniform on $I_Z \times I_W$, and take $\lambda_{j,PC}$ as the proportion of all of those $U_1,\ldots,U_{N\cdot n}$ closest to (Z_j,W_j), multiplied by $\operatorname{Area}(I_Z \times I_W)$. This requires $N \cdot n^2$ comparisons. The resulting estimator of the scaling constant is

$$\hat{a}_{I_X \times I_W,n} = \sum_{j=1}^{n} v\big(X_j,W_j\big)\,\lambda_{j,PC}. \tag{27}$$

The contrast is then constructed as before, see the supplement for further details. We only give the formula for $B \sim N(\mu,\Sigma)$ multivariate normal, which is used in the simulations and application. Given

intervals $I_{1,X},\dots,I_{p,X}\subset I_X$ and $I_{1,W},\dots,I_{q,W}\subset I_W$ and an $s>0$, we need to minimize

$$\begin{aligned}M(\mu,\Sigma)=\sum_{k=1}^{p}\sum_{l=1}^{q}\Big(&\hat{a}_n\int_{I_{k,X}\times I_{k,X}}\int_{I_{l,W}\times I_{l,W}}\varphi\big(0;\mu_1(x_2-x_1)+\mu_2(w_2-w_1),\big((1,x_1,w_1)\Sigma(1,x_1,w_1)'\\&+(1,x_2,w_2)\Sigma(1,x_2,w_2)'^{-2}\big)^{-1}\big)\,dx_1\,dx_2\,dw_1\,dw_2\\-2\sum_{j=1}^{n}&\int_{I_{k,X}}\int_{I_{l,W}}\varphi\big(0;Y_j-\mu_0-\mu_1x-\mu_2w,\big((1,x,w)\Sigma(1,x,w)'^{-2}\big)^{-1}\big)\\&1_{I_{k,X}}(X_j)\,1_{I_{l,W}}(W_j)\,\lambda_{j,PC}\,dx\,dw\Big),\end{aligned}$$

where $\hat{a}_n=\hat{a}_{I_X\times I_W,n}$.

5. Simulation Study

We investigate the finite-sample performance of the semiparametric estimators of Section 4 in a simulation study.

5.1. Simulation in the model (2)

Data generating process.

- A_1 independent of (A_0,B_0,B_1)
- $A_1\sim 0.5\cdot\text{Beta}(2,2)+1$, where $\text{Beta}(\alpha,\beta)$ is the beta-distribution with parameters α and β,
- $A_0\sim U(0,3)$, and

$$(B_0,B_1)'\sim N\big(\mu,\Sigma\big),\quad \mu=(5,2)',\qquad \Sigma=\begin{pmatrix}4 & 1.4\\ 1.4 & 1\end{pmatrix},$$

- The joint dependence of (A_0,B_0,B_1) is given by a Gaussian copula with parameters $\rho_{B_0,A_0}=\rho_{B_1,A_0}=0.9$ (and of course $\rho_{B_0,B_1}=0.7$).
- $Z\sim 12\ \cdot\text{Beta}(0.5,0.5)$.

For this data-generating process, the maximal interval that satisfies (20) is $I=[3,12]$, see the example following (20). The proportion of X-values that falls into I is about $44,3\%$ (large-scale simulation), thus the effective sample size is 0.443 times the actual sample size when using the full interval I.

Estimation of the scaling constant. First, we consider the estimator of the scaling constant $E|A_1|^{-1}$ from the first-stage equation $X=A_0+A_1Z$. Simulating directly a sample of size $n=10^8$ for A_1 and estimating $E|A_1|^{-1}$ by the mean of the inverses gives ≈ 0.8065. The results for the estimator (21) for various sample sizes, where v is taken as uniform weight $1/9$ over the interval I, are presented in Table (1). One can, in particular, see the parametric rate of convergence quite clearly.

n	Bias	Std	MSE	Total SE	Rel. Error
$5\cdot 10^2$	0.0013	0.027	$7.3\cdot 10^{-4}$	0.36	0.033
$1\cdot 10^3$	$8\cdot 10^{-4}$	0.018	$3.5\cdot 10^{-4}$	0.35	0.023
$5\cdot 10^3$	$1\cdot 10^{-4}$	0.008	$0.7\cdot 10^{-4}$	0.36	0.010
$10\cdot 10^3$	$0.5\cdot 10^{-4}$	0.006	$0.36\cdot 10^{-4}$	0.36	0.007
$20\cdot 10^3$	$0.2\cdot 10^{-4}$	0.004	$0.18\cdot 10^{-4}$	0.36	0.005
$50\cdot 10^3$	$0.1\cdot 10^{-4}$	0.002	$0.07\cdot 10^{-4}$	0.36	0.003

Table 1: Statistics for the scaling constant

Estimating the parametric model. Next, we consider estimation of the parameters of the normal distribution of (B_0, B_1).

In the contrast function (23), we choose the weighting measure $d\nu(t)$ to be centered Gaussian with standard deviation $s = 0.1$, resulting in $M(\mu,\Sigma)$ as given in (25). We partition the interval I into $q = 10$ successive equal-length subintervals I_p, $p = 1,\dots,10$. The integrals are computed numerically using the function `adaptIntegrate` contained in the R-library `cubature`.

For illustration, we first consider a single sample of size 5000. If we directly compute the correlation between the B_j's and X, we find about 0.125 for both $j = 0,1$, so that there is some endogeneity in the model. A simple least-squares fit of X on Y gives the coefficients $\hat{b}_0 = 3.4$ and $\hat{b}_0 = 2.3$, so that in particular the mean of the intercept is estimated incorrectly which is well in line with our theory.

Now, a 5-dimensional grid search on a large grid is evidently computationally infeasible, at least in repeated simulations. Therefore, we first compute an IV-fit using the R-library `AER` and the function `ivreg`. For the sample above, the estimates for the coefficients are given by $\hat{b}_0 = 5.91$ and $\hat{b}_1 = 1.98$.

Since from Section 2.3, $\hat{b}_1 = 1.98$ is consistent for μ_{B_1}, we take $\hat{\mu}_{B_1} = \hat{b}_1 = 1.98$.

In a next step, we fix the values of the means as those of the IV fit, and minimize the criterion (25) with respect to the parameters of Σ by using the numerical routine `nlm`, an implementation of the Nelder-Mead algorithm. Here, the covariance matrix is parametrized by using its Cholesky decomposition. As starting values for the variances we take the rescaled fit for the estimated coefficients in the IV-regression, except for the covariance which is set to zero. In the present sample, this is $\mathrm{diag}(3.77, 0.04)$. In particular in the variance of B_1, this is way of the true value.

In terms of standard deviations and correlation, the resulting estimates are $\hat{\sigma}_{B_0} = 2.02$ (true = 2), $\hat{\sigma}_{B_1} = 0.97$ (true = 1) and $\hat{\rho}_{B_0,B_1} = 0.66$ (true = 0.7).

Finally, we fix the above estimates of $\hat{\sigma}_{B_j}$, $j = 0,1$, $\hat{\rho}_{B_0,B_1}$ and $\hat{\mu}_{B_1}$, and determine the estimate of μ_{B_0} by using a grid search of criterion (25). Here, we use a grid of width 0.1 from $\hat{b}_0 - 1$ to $\hat{b}_0 + 1$. The resulting estimate in this sample is 5.01.

We now perform this algorithm for various sample sizes repeatedly, the results can be found in Tables 2 and 3. The estimates of all parameters are quite reasonable. The MSE for estimating μ_0 is much higher than for μ_1, which is well in line with our theory, which shows that also identification of μ_0 is much harder (and weaker) than of μ_1. The estimates of the parameters of the covariance matrix are also acceptable, although σ_0 and ρ seem to have a small bias.

N		Mean	Bias	Std	MSE	Total SE	Rel. Error
2000	μ_0	5.058	0.058	0.482	0.235	471	0.097
	μ_1	1.999	-0.001	0.053	0.003	6	0.026
5000	μ_0	5.017	0.017	0.334	0.112	560	0.067
	μ_1	1.999	-0.001	0.035	0.001	6	0.017
10000	μ_0	4.993	-0.007	0.229	0.053	526	0.046
	μ_1	2.000	0.000	0.024	0.001	6	0.012
20000	μ_0	5.001	0.001	0.163	0.027	532	0.033
	μ_1	2.000	0.000	0.016	0.000	5	0.008

Table 2: Statistics for the coefficients μ_0 and μ_1

N		Mean	Bias	Std	MSE	Total SE	Rel. Error
2000	σ_0	2.048	0.048	0.034	0.003	0.030	7
	σ_1	1.009	0.009	0.131	0.017	0.131	34
	ρ	0.647	-0.053	0.156	0.027	0.236	54
5000	σ_0	2.047	0.047	0.025	0.003	0.027	14
	σ_1	0.998	-0.002	0.093	0.009	0.093	43
	ρ	0.650	-0.050	0.142	0.023	0.215	114
10000	σ_0	2.049	0.049	0.013	0.003	0.026	26
	σ_1	1.002	0.002	0.071	0.005	0.071	50
	ρ	0.664	-0.036	0.009	0.001	0.053	14
20000	σ_0	2.050	0.050	0.012	0.003	0.026	53
	σ_1	0.996	-0.004	0.063	0.004	0.063	79
	ρ	0.662	-0.038	0.048	0.004	0.088	75

Table 3: Statistics for the coefficients σ_0,σ_1 and ρ

5.2. Simulation in the extended model (1)

Data generating process and estimates of scaling constant.

- A_1 independent of (A_0, A_2, B'),
- $A_1 \sim 0.5 \cdot \text{Beta}(2,2) + 1$,
- $A_0 \sim U(0,3)$, and $A_2 \sim 0.2 \cdot \text{Beta}(2,2)$.

$$B = \begin{pmatrix} B_0 \\ B_1 \\ B_2 \end{pmatrix} \sim N(\mu, \Sigma), \quad \mu = (5,2,3)', \qquad \Sigma = \begin{pmatrix} 4 & 0.8 & 0.9 \\ 0.8 & 1 & 0.9 \\ 0.9 & 0.9 & 2.25 \end{pmatrix},$$

- The joint dependence of (A_0, A_2, B') is given by a Gaussian copula with correlation parameters $\rho_{B_0,A_0} = \rho_{B_2,A_2} = 0.7$, $\rho_{B_1,A_0} = \rho_{B_2,A_0} = \rho_{B_0,A_2} = \rho_{B_1,A_2} = 0.5$ and $\rho_{A_0,A_2} = 0.2$ (and of course $\rho_{B_0,B_1} = 0.4$, $\rho_{B_0,B_2} = 0.3$, $\rho_{B_1,B_2} = 0.6$).
- $Z \sim 15 \cdot \text{Beta}(0.9, 0.9)$, $W \sim 10 \cdot \text{Beta}(0.7, 0.7)$, and their dependence is determined by a Gaussian copula with correlation $\rho_{Z,W} = 0.7$.

For this data-generating process, the support condition (26) is satisfied for $I_X = [5,15]$ and $I_W = [0,10]$. The proportion of X-values that falls into I_X is about 54% (large-scale simulation, I_W is the full support of W).

We use the simulated Priestly-Chao weights as proposed in Section **??** for $N = 200$ (an increase to $N = 500$ did not improve the succeeding estimates, as we demonstrate in Table 4 on the scaling constant). Though computationally intensive, note that these need to be computed only once, and can be used both for estimation of the scaling constant as well as for computing the semiparametric estimator. Computing the weights once for $n = 20000$ take ≈ 3 min on a CORE I7 computer.

Table 4 contains the simulations results for the estimator of the scale constant including the W.

	Mean	Std. Dev	Abs. Bias	MSE	Mean Sq. Bias	Total SE	Rel. Error
n = 2000, N = 200	0.812	0.017	0.014	0.001	0.00032	1.24	0.031
n = 5000, N = 200	0.809	0.011	0.009	0.000	0.00014	1.34	0.020
n = 10000, N = 200	0.808	0.008	0.006	0.000	0.00007	1.31	0.014
n = 20000, N = 200	0.807	0.006	0.004	0.000	0.00003	1.23	0.010
n = 2000, N = 500	0.812	0.017	0.014	0.001	0.00034	1.27	0.031
n = 5000, N = 500	0.809	0.012	0.009	0.000	0.00014	1.36	0.020
n = 10000, N = 500	0.808	0.008	0.006	0.000	0.00006	1.27	0.014
n = 20000, N = 500	0.807	0.006	0.004	0.000	0.00003	1.23	0.010

Table 4: Statistics for the scaling constant

Estimating the parametric model. In order to investigate the endogeneity in the model, we first simulate a very large sample of size 10^7, and compute correlations of X with the coefficients B_j, $j = 0,1,2$, which are all about 0.1. The OLS estimates are strongly biased, while IV-estimates give $\hat{b}_0 = 5.44$, $\hat{b}_1 = 2.00$ and $\hat{b}_2 = 3.03$. Thus, the bias in $\hat{b}_0$ is clearly visible, while the bias in $\hat{b}_2$ is small due to small variation in A_2, but significant (standard deviation estimated at $2 \cdot 10^{-7}$). Nevertheless, we can only expect to improve upon $\hat{b}_2$ in very large samples.

Now, the further estimation algorithm for the coefficient μ and Σ is completely analogous to that in the previous subsection, we only perform a grid search w.r.t. both μ_0 and μ_2. The results for sample sizes $n = 2000$ and $n = 10000$ are contained in Tables 5 - 6. Here we use $N = 200$, increasing this to $N = 500$ does not change the results much. The estimates of the parameters of the covariance matrix are slightly biased, but still reasonable.

	Mean	Std. Dev	Mean Abs. Bias	MSE	Mean Sq. Bias	Total SE	Rel. Error
mu0	5.0885	0.8323	0.6728	1.393335	0.700582	2787	0.236
mu1	1.9993	0.1053	0.0838	0.022176	0.011088	44	0.074
mu2	3.0067	0.2584	0.2109	0.133556	0.066800	267	0.122
sig0	2.2003	0.1362	0.2159	0.077237	0.058677	154	0.139
sig1	1.0999	0.2592	0.2233	0.144339	0.077160	289	0.380
sig2	1.3695	0.5532	0.4622	0.629084	0.323064	1258	0.529
rho01	0.3754	0.2999	0.2016	0.180478	0.090542	361	1.062
rho02	0.3338	0.1680	0.1158	0.057624	0.029383	115	0.800
rho12	0.4931	0.2790	0.1546	0.167135	0.089283	334	0.681

Table 5: Statistics for sample size $n = 2000$

	Mean	Std. Dev	Mean Abs. Bias	MSE	Mean Sq. Bias	Total SE	Rel. Error
mu0	5.0369	0.4826	0.3971	0.467200	0.234279	4672	0.137
mu1	1.9990	0.0483	0.0387	0.004670	0.002335	47	0.034
mu2	2.9964	0.1681	0.1415	0.056520	0.028267	565	0.079
sig0	2.1849	0.0701	0.1853	0.044016	0.039104	440	0.105
sig1	1.0243	0.1436	0.1200	0.041826	0.021208	418	0.205
sig2	1.3634	0.3025	0.2745	0.201634	0.110150	2016	0.299
rho01	0.3959	0.1401	0.1075	0.039272	0.019644	393	0.495
rho02	0.3832	0.0519	0.0882	0.012302	0.009613	123	0.370
rho12	0.6064	0.0747	0.0545	0.011193	0.005617	112	0.176

Table 6: Statistics for sample size $n = 10000$

6. Application

6.1. Motivation: Consumer Demand

Both heterogeneity and endogeneity play an important role in classical consumer demand. The most popular class of parametric demand systems is the almost ideal (AI) class, pioneered by Deaton and Muellbauer (1980). In the AI model, instead of quantities budget shares are being considered, and they are being explained by log prices and log total expenditure[4]. The model is linear in log prices and a term that involves log total expenditure linearly, but divided by a price index that depends on parameters of the utility function. In applications, one frequent shortcut is that the price index is replaced by an actual price index, another is that homogeneity of degree zero is imposed, which means that all prices and total expenditure are relative to a price index. This step has the beneficial side effect that it removes general inflation as well. A popular extension in this model allows for quadratic terms in total expenditure (QUAIDS, Banks, Blundell and Lewbel (1997)). However, we focus on the budget share for food at home (BSF), which, due at least in parts to satiation effects, is often documented to decline steadily across the total expenditure range. This motivates our individual level specification $BSF_i = b_{0i} + b_{1i}\ln(TotExp_i) + b_{2i}\ln(Foodprice_i)$, where $TotExp_i$ and $Foodprice_i$ are the variables as described above. To relate it to the population model, we allow now for the intercept b_{0i} to be a deterministic function of observable demographic variables W_i and a time variable T_i as well, and for all coeffcients b_i to in addition vary across the populations, leading the overall model

$$BSF_i = B_{0i} + B_{1i}\ln(TotExp_i) + B_{2i}\ln(Foodprice_i) + b_3W_{1i} + b_4W_{2i} + b_5T_i,$$

As mentioned, frequently endogeneity of total expenditure is being assumed (see Blundell, Pashardes and Weber (1995), Lewbel (1999)), in parts because food expenditure accounts for a large fraction of total expenditure and an IV approach advocated. In our setup, this equation takes the form

$$\ln(TotExp_i) = A_{0i} + A_{1i}\ln(Income_i) + A_{2i}\ln(Foodprice_i) + a_3W_{1i} + a_4W_{2i} + a_5T_i,$$

and the standard argument for the validity of income is that for the type of households we consider (two person households, no children), labor supply is rather inelastic and variations in labor income are hence largely a function of variations in the wage rate, which is plausibly exogenous. Note that we include the price of food as exogenous regressor, as variations in this variable cover some of the exogenous variation

[4]The use of total expenditure as wealth concept is standard practise in the demand literature and, assuming the existence of preferences, is satisfied under an assumption of separability of the labor supply from the consumer demand decision, see Lewbel (1999).

in food expenditure, which in turn account for some of the endogeneity in total expenditure. We also control again for observables, including a time trend.

6.2. The Data: The British Family Expenditure Survey

The FES reports a yearly cross section of labor income, expenditures, demographic composition, and other characteristics of about 7,000 households. We use the years 1994-2000, but exclude the respective Christmas periods as they contain too much irregular behavior. As is standard in the demand system literature, we focus on the subpopulation of two person households where both are adults, at least one is working, and the head of household is a white collar worker. This is to reduce the impact of measurement error; see Lewbel (1999) for a discussion.

We form several expenditure categories, but focus on the food at home category. This category contains all food expenditure spent for consumption at home; it is broad since more detailed accounts suffer from infrequent purchases (the recording period is 14 days) and are thus often underreported. Food consumption accounts for roughly 20% of total expenditure. Results actually displayed were generated by considering consumption of food versus nonfood items. We removed outliers by excluding the upper and lower 2.5% of the population in the three groups. We form food budget shares by dividing the expenditures for all food items by total expenditures, as is standard in consumer demand.

To obtain the respective own relative prices, we normalize price by dividing by the general price index excluding food (i.e., we consider the price of food vs. the price of all nondurable goods except food). We also divide total expenditure by the price index. As already mentioned, we use labor income as an instrument. Labor income is constructed as in the household below average income study (HBAI), i.e., it is roughly defined as labor income after taxes and transfers. We include the remaining household covariates as regressors. Specifically, we use principal components to reduce the vector of remaining household characteristics to a few orthogonal, approximately continuous components, mainly because we require continuous covariates for estimation. Since we already condition on a lot of household information by using the specific subgroup, we only use two principal components, denoted W_{1i} amd W_{2i}. While this is arguably ad hoc, we perform some robustness checks like alternating the component or adding several others, and the results do not change appreciably. Finally, we also use a monthly time trend, denoted T_i. The following table provides some descriptive statistics of the data:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St. Dev.
Food share	0.0023	0.1419	0.1982	0.2154	0.2741	0.7840	0.1020
ln Food prices	-0.1170	-0.0049	0.0125	0.0173	0.0408	0.1748	0.1492
ln Expenditures	3.2940	4.6290	5.2810	5.2050	5.8040	6.9270	0.1280
ln Income	3.5040	4.7350	5.3490	5.2960	5.8790	6.9310	0.7771
PC1	-1.8810	-0.9552	0.1081	0.0038	0.8620	2.0030	0.9885
PC2	-2.9070	-0.7878	-0.1426	0.0174	0.9447	2.2540	0.9991

6.3. Details of the Econometric Implementation

We outline now our estimation strategy. The model we are estimating is as displayed in Section 6.1. We first use the IVREG function of the AER package in R to run the regression above, then subtract the terms involving the deterministic coefficients, i.e., we form $\widetilde{BSF}_i = BSF_i - \left[\widehat{b}_3 W_{1i} + \widehat{b}_4 W_{2i} + \widehat{b}_5 T_i\right]$, and

	Estimate	Std. Error	t-value	$Pr(>\|t\|)$
E(B_0)	0.6330	0.0072	87.3871	$< 2.2 \cdot 10^{-16}$***
E(B_1)	-0.0998	0.0014	-72.1405	$< 2.2 \cdot 10^{-16}$***
E(B_2)	-0.0842	0.0283	-2.9716	0.0028**

Table 7: IV coefficients. Significance codes: ***, 0.001; **, 0.01.

$\ln(\widetilde{TotExp_i}) = \ln(TotExp_i) - [\widehat{a}_3 W_{1i} + \widehat{a}_4 W_{2i} + \widehat{a}_5 T_i]$, where the hats denote IV estimates. This is justified, because as we have shown above, IV produces consistent estimates for a fixed coefficient, provided the first stage has no randomness. These variables become our new dependent variables, resulting in the specification:

$$\begin{aligned} \widetilde{BSF}_i &= B_{0i} + B_{1i}\ln(TotExp_i) + B_{2i}\ln(Foodprice_i), \\ \ln(\widetilde{TotExp_i}) &= A_{0i} + A_{1i}\ln(Income_i) + A_{2i}\ln(Foodprice_i). \end{aligned}$$

This model is apparently of our extended type, with $X_i = \ln(TotExp_i)$, $Z_i = \ln(Income_i)$, and $W_{0i} = \ln(Foodprice_i)$.

As interval of X_i used for estimation we take $[4.3, 6.1]$.

To optimize the criterion function, we separate the parameter space into two parts, we first optimize over the covariances by fixing the means and applying a gradient-based algorithm, then optimize over the means by searching over a grid.

These alternating steps were iterated up to three times to ensure convergence, using the new means and covariances as starting values. However, there was no change in the optimal parameters after the first iteration, up to computation error.

Our IV starting values are presented in the next table, but recall that these values, in particular the estimator for EB_2, may be biased. Indeed, that is what we will find out in the next subsection.

More details about the R implementation of the optimization step follow.

To find the minimizer of our objective function, we use the NLM function of the R package with these initial specifications: p: initial value $diag\{.6, .02, .02\}$, $gradtol$: minimum value of scaled gradient 10^{-9}, $steptol$: minimum allowable relative step length. 10^{-9}. In our application, the results appear to be somewhat sensitive to the choice of these values, but only in as much as a wrong choice will either lead to explosive results that are obviously unreasonable, or cause the optimizer to stall after zero iterations.

The next step in the optimization is to recompute the means by minimizing the objective function over the mean parameter. This time we simply search over 25-point grids covering the interval

$$[\hat{E}B_j - 5 * |\hat{E}B_j|, \hat{E}B_j + 5 * |\hat{E}B_j|]$$

where $\hat{E}B_j$ is the estimate of the mean computed in the previous iteration, or the IV estimates for the very first iteration.

Finally, we repeat the iteration up to three times, again without any appreciable change up to numerical error.

6.4. Results

To analyze our main results, we display them in two different ways. First, we show a series of graphs individual mean and variance parameter estimates with associated bootstrapped distribution statistics; second, we show the resulting densities. The point estimates for the mean parameters are given in the table below:

	Point Estimate	Mean	Median	Std. Dev.	.275 Quantile	.975 Quantile
E(B_0)	0.6331	0.6338	0.6334	0.0199	0.5975	0.6706
E(B_1)	-0.0999	-0.1002	-0.0998	0.0037	-0.1075	-0.0934
E(B_2)	-0.2598	-0.2546	-0.2593	0.1102	-0.4235	-0.0103

As we see, the estimates for the mean are generally precisely estimated. The mean coefficients are of a very sensible magnitude. Given that log Total Expenditure varies roughly between 3 and 6, with every near tripling of income we observe a decrease in the food demand budget share by 10 percentage points, say, from 27 to 17 percent. Also, since prices are measured in relative units, a relative price of 1.07 corresponds approximately to a log price of 0.07. Thus, an increase in the relative price of 7%, from 1 to 1.07 corresponds roughly to a decrease of the food budget share by 1.7 percent, say, the budget share drops in response from 25.8 to 24.1 percent. Since prices generally vary between -0.11 and 0.17, this means that the budget share of a person with average price semi-elasticity is, at least in our data, not strongly affected by the historical changes in the relative price found in our data.

The comparison between the IV starting values and the final values of the mean coefficients is quite informative, and generally confirms with theory: Corresponding to the fact that the IV estimate of EB_1 is known to be unbiased while the one for EB_2 is not, we find very little movement in the former coefficient, while the second nearly triples. In fact, price effects only become significantly negative after applying our procedure, lending credibility to our approach and also emphasizing the role of the bias, if unobserved heterogeneity is not appropriately modelled. The variance parameters are generally less precisely estimated, in particular, there seem to more mass in the tails of the bootstrap distribution. There is not a lot of evidence of covariance between the random slopes. The variance is sizeable relative to the magnitude of the mean effects, implying that the average effects mask profound heterogeneity.

	Point Estimate	Standard Error
Var(B_0)	0.4712	0.0641
Var(B_1)	0.0153	0.0225
Var(B_2)	0.0180	0.0066
Cov(B_0, B_1)	-0.0850	0.1050
Cov(B_0, B_2)	-0.0297	0.0276
Cov(B_1, B_2)	0.0053	0.0064

To interpret this type of heterogeneity, it is advantageous to display the resulting random coefficients density. We focus on the results concerning the slope parameters, they are given in figures 2, 3 and 4, below.

As is obvious, most of the individuals reduce their budget share of food as total expenditure and prices increase. There is pronounced heterogeneity when it comes to degree of reduction. Indeed, especially with total expenditures, some individuals even respond with increases in their budget share, however, only very light ones. One issue in that respect is whether the parts of the population density which correspond to positive effects is statistically significant from zero. While we leave the derivation of such a test, which would extend standard nonparametric hypothesis tests for densities, for future research, given the pronounced uncertainty associated with the variance parameters we feel that it is questionable

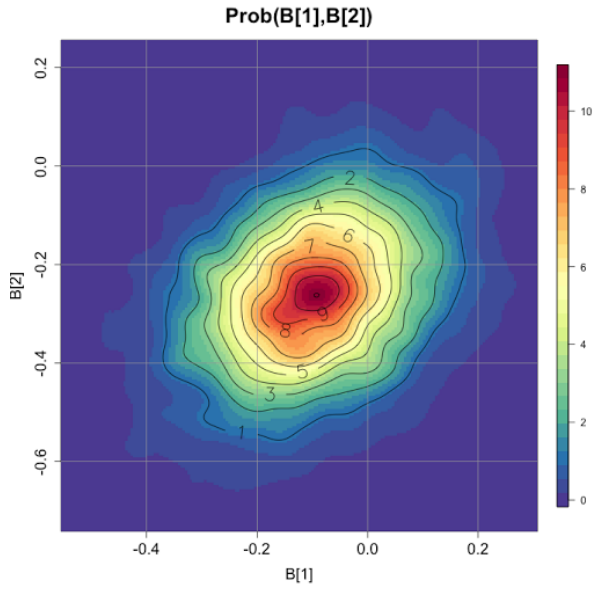

Figure 2: Marginal distribution of the slope on log total expenditures (B_1) and log food prices (B_2) when the IV interval length is 1.600.

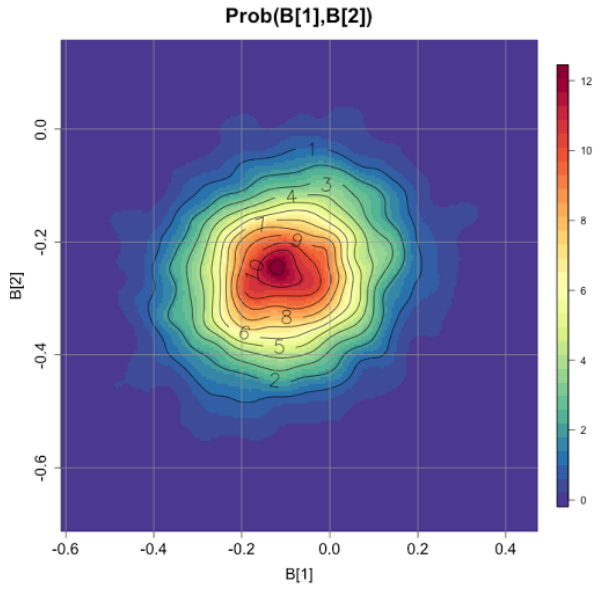

Figure 3: Marginal distribution of the slope on log total expenditures (B_1) and log food prices (B_2) when IV interval length is 1.650.

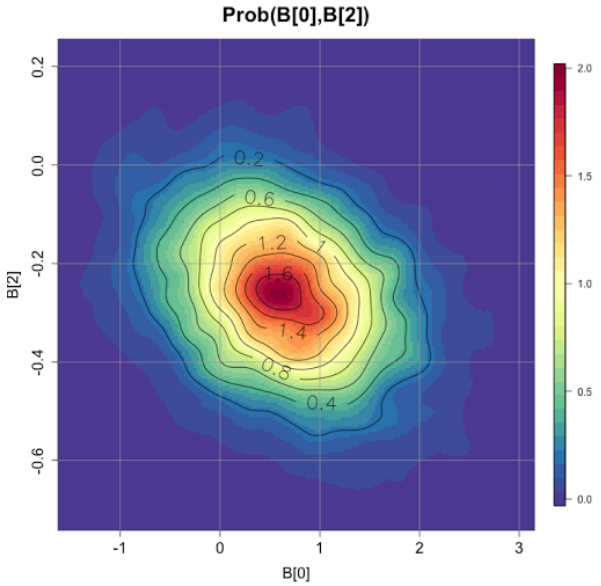

Figure 4: Marginal distribution of the intercept (B_0) and log food prices (B_2).

whether there are any individuals whose marginal total expenditure effects are beyond 0.1. Having said that, we believe that it is entirely possible that some individuals have small positive effects, in particular those at the lower end of the total expenditure distribution (recall that total expenditure and preferences may well be correlated). Note moreover that our functional form assumption (normality) forces the density to be roughly symmetric.[5] In general, however, the result along the total expenditure dimension are very plausible, including the magnitude of such effects.

This is the more true for price effects. Note that virtually the entire population responds to a food price increase with a decrease in the budget share of this good. While some individuals reduce their food demand only very lightly, as is evident by values of the semi-elasticity of - 0.1, which translate into below 1% reductions in budget shares for a 7% increase in price, others respond much more strongly with a large fraction of the population having values of around -0.4, corresponding to a 3-4% reduction for the same relative price change. This suggests that between the 25% least and most reactive individuals there is roughly a four fold difference in the strength of their price effects. This has strong implications for welfare analysis, as the welfare effects are largely built on both coefficients.

In sum, the application reveals that our method is able to remove biases stemming from the omission of unobserved heterogeneity. It is also able to capture the heterogeneity in a, as we feel, concise and practical fashion. Since the purpose of this application is more illustrative, we refer the interested reader to the authors' website for more details of the application.

7. Conclusion

This paper analyzed the triangular model with random coefficients in the first stage and the outcome equation. We show that in this class of models, the joint distribution of parameters, as well as important marginal densities are generically not point identified, even if the instruments enter monotonically. Based on these results, we provide additional restrictions that ensure point identification of the marginal distribution of parameters in the outcome equation. These restrictions are for instance satisfied, if one of

[5] We leave a more thorough examination of this issue, including the use of left skewed distributions, for future research.

the coefficients in the first stage equation is nonrandom. We establish that even in the presence of these restrictions, standard linear IV does not produce consistent estimates of the average effects. This motivates our search for a (semi)parametric estimator that is relevant for applications, and incorporates the conclusions we draw from the (non-)identification results. An alternative strategy is to follow a partial identification approach. While this paper briefly discusses such an approach, it leaves further details for future research.

References

[1] BANKS, J., R. BLUNDELL, and A. LEWBEL, 1997. "Quadratic Engel Curves And Consumer Demand," Review of Economics and Statistics, vol. 79(4), 527-539.

[2] BERAN, R. and P. HALL (1992): Estimating Coefficient Distributions in Random Coefficient Regressions," *Annals of Statistics*, 1970-1984

[3] BERAN, R., A. FEUERVERGER, and P. HALL (1996): "On Nonparametric Estimation of Intercept and Slope in Random Coefficients Regression". *Annals of Statistics*, **2**, 2569–2592.

[4] BLUNDELL, R., KRISTENSEN, D., and MATZKIN, R. (2011); Stochastic Demand and Revealed Preference, *Journal of Econometrics*, forthcoming.

[5] BONHOMME, A. (2012). Identifying Distributional Characteristics in Random Coefficients Panel Data Models. *Review of Economic Studies*, **79**, 987-1020

[6] CHAMBERLAIN, G. (1982), "Multivariate Regression Models for Panel Data," *Journal of Econometrics,* 18(1), 5 - 46.

[7] CHAMBERLAIN, G. (1992): Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567-596.

[8] CHESHER, A. (2003) "Identification in Nonseparable Models", *Econometrica*, 71, 1405-1441.

[9] CHESHER, A. and A. ROSEN (2013). Generalized instrumental variable models, CeMMAP working papers CWP04/14, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

[10] DEATON, A. and J. MUELLBAUER (1980). An Almost Ideal Demand System, *American Economic Review*, 70, 312-326

[11] FLORENS, J.P., J.J. HECKMAN, C. MEGHIR, and E. VYTLACIL (2008) "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, Vol. 76, No. 5, pp. 1191-1206

[12] FOX, J. and A. GANDHI, (2009): "Identifying Heterogenity in Economic Choice and Selection Models Using Mixtures". Working Paper.

[13] GAUTIER, E. and S. HODERLEIN (2012): "Estimating the Distribution of Treatment Effects". CeMMAP Working Paper.

[14] GAUTIER, E. and Y. KITAMURA (2013). Nonparametric Estimation in Random Coefficients Binary Choice Models. *Econometrica*, **81**, 581-607.

[15] D'HAULTFOEUILLE, X., HODERLEIN, S. and SASAKI (2014). Nonlinear Difference-in-Differences in Repeated Cross Sections with Continuous Treatments, Working Paper, Boston College.

[16] HAUSMAN, J. and W. NEWEY, (2011),Individual Heterogeneity and Average Welfare, *Working Paper*, MIT.

[17] HECKMAN, J. and E. 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," *The Journal of Human Resources*, 33, 974 - 987. 2, 9, 19

[18] HODERLEIN, S, J. KLEMELÄ, and E. MAMMEN (2010): "Analyzing the Random Coefficient Model Nonparametrically". *Econometric Theory*, **26**, 804–837.

[19] HODERLEIN, S. and E. MAMMEN (2007); "Identification of Marginal Effects in Nonseparable Models without Monotonicity, *Econometrica,* 75, 1513–1518.

[20] HODERLEIN, S. and B. SHERMAN (2012): Identification and Estimation in a Correlated Random Coefficients Binary Response Model", CeMMAP Working Paper.

[21] IMBENS, G. and W. NEWEY (2009); Identification and Estimation of Triangular Simultaneous Equations Models without Additivity, *Econometrica*, forthcoming.

[22] KASY, M. (2011) "Identification in Triangular Systems using Control Functions," *Econometric Theory*, **27**, 663–671.

[23] KASY, M. (2013) "Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment," *Review of Economic Studies*, forthcoming.

[24] LEWBEL, A. (1999); "Consumer Demand Systems and Household Expenditure", in Pesaran, H. and M. Wickens (Eds.), Handbook of Applied Econometrics, Blackwell Handbooks in economics.

[25] LEWBEL, A. and K. PENDAKUR (2014); "Unobserved Preference Heterogeneity in Demand Using Generalized Random Coefficients, Working Paper."

[26] MASTEN, M. (2013); "Random Coefficients on Endogenous Variables in Simultaneous Equation Models", Working Paper, Duke University.

[27] MASTEN, M. and A. TORGOVITSKY (2014); "Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model", Working Paper.

[28] MATZKIN, R. (2007);"Nonparametric Identification" in J.J. Heckman and E.E. Leamer (ed.) *Handbook of Econometrics*, Vol. 6, Ch. 73.

[29] MATZKIN, R. (2012); "Identification in nonparametric limited dependent variable models with simultaneity and unobserved heterogeneity," *Journal of Econometrics*, 166, 106-115.

[30] WOOLDRIDGE, J. (1997): "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models," *Economics Letters*, 56, 129-133.

A. Identification / Nonidentification: Proofs

A.1. Proofs for Section 2

Proof of Lemma 2.1. Since the distribution of Z is fixed, it suffices that the conditional distributions of $(Y,X)|Z=z$ and $(\tilde{Y},\tilde{X})|Z=z$ coincide for all $z \in \text{supp } Z$, or, equivalently, that the conditional characteristic functions coincide, which immediately follows from (5) and the assumption in the lemma. ∎

Proof of Theorem 1. In model (2) with reduced form (3), we denote by $u = (u_0, u_1)'$ (respectively $v = (v_0, v_1)'$) the coordinates corresponding to $A = (A_0, A_1)'$ (respectively $C = (C_0, C_1)'$) in Fourier space. Further, we write (u, v) and (a, c), $a = (a_0, a_1)'$, $c = (c_0, c_1)'$, instead of $(u', v')'$ or $(a', c')'$.

Step 1. In the first step, we construct two appropriate joint densities of (A, C), whose characteristic functions coincide on the set $\mathscr{S}$ in Lemma 2.1.

We introduce the density on the line

$$g_\beta(s) = \alpha\beta \cdot \exp\big(1/(\beta^2 s^2 - 1)\big)\, 1_{(-1,1)}(\beta s), \quad s \in \mathbb{R}, \tag{28}$$

where $\alpha > 0$ is a normalizing constant and $\beta > 0$ is the scale parameter. Note that g_β is supported on $[-1/\beta, 1/\beta]$ and differentiable infinitely often on the whole real line. Consider the product density

$$G_\beta(a, c) = \prod_{j=0}^{1} g_\beta(a_j) g_\beta(c_j), \qquad (a, c) \in \mathbb{R}^4,$$

which is supported on $[-1/\beta, 1/\beta]^4$ and differentiable infinitely often.

Consider the non-constant polynomial $Q(u, v) = u_0 v_1 - u_1 v_0$, $(u, v) \in \mathbb{R}^4$, and recall that for the set $\mathscr{S}$ in Lemma 2.1,

$$\mathscr{S} \subseteq \{(u, v) \in \mathbb{R}^4 : \, Q\big(-i(u, v)\big) = 0\}.$$

Set

$$r(a, c) = \big[\big(\partial_{a_0}\partial_{c_1} - \partial_{a_1}\partial_{c_0}\big) G_1\big](a, c) =: \big[Q\big(\partial_{a_0}, \partial_{a_1}, \partial_{c_0}, \partial_{c_1}\big) G_1\big](a, c)$$

Since the Fourier transform turns differentiation into multiplication, we obtain

$$(\mathscr{F}_4 r)(u, v) = Q\big(-i(u, v)\big)\, (\mathscr{F}_4 G_1)(u, v), \qquad (u, v) \in \mathbb{R}^4,$$

which vanishes on $\mathscr{S}$. We therefore set

$$\tilde{f}_j(u, v) = G_{1/2}(u, v) + (-1)^j \gamma \cdot r(u, v), \quad j = 1, 2, \quad (u, v) \in \mathbb{R}^4,$$

where the constant $\gamma > 0$ is (and can be) chosen such that $\tilde{f}_j$ are non-negative functions. These are differentiable infinitely often, their support is included in $[-2, 2]^4$, and their Fourier transforms turn out to be

$$\big(\mathscr{F}_4 \tilde{f}_j\big)(u, v) = \big(\mathscr{F}_4 G_{1/2}\big)(u, v) + (-1)^j \gamma \cdot Q\big(-i(u, v)\big)\, \big(\mathscr{F}_4 G_1\big)(u, v). \tag{29}$$

Since $Q(0) = 0$ the functions $\tilde{f}_j$ integrate to one so that they are (two distinct) probability densities.

Step 2. The second step consists of a simple translation, which allows to relate to the density of (A, B).

Define the shifted densities $f_j(a, c) = \tilde{f}_j\big((a, c) - (0, 3, 0, 0)\big)$, so that

$$\big(\mathscr{F}_4 f_j\big)(u, v) = \exp(3iu_1)\big(\mathscr{F}_4 \tilde{f}_j\big)(u, v). \tag{30}$$

and

$$\int_{\mathbb{R}^4} 1_{(-\infty, 1]}(a_1) f_j(a, c)\, da\, dc = 0.$$

Suppose that (A, C) and $(\tilde{A}, \tilde{C})$ have the densities f_1 and f_2, respectively. Then their characteristic functions coincide on the set $\mathscr{S}$, and hence the joint densities of the observed variables also coincide by Lemma 2.1. Furthermore, both $A_1 > 1$ and $\tilde{A}_1 > 1$ a.s., and by the change of variables formula (4) the corresponding (distinct) densities $f_{A,B}$ and $f_{\tilde{A},\tilde{B}}$ will also be differentiable infinite often and of compact support.

At this stage, we have already shown non-identifiability of the joint distribution of (A, B) in the class of distributions with infinitely differentiable densities with compact support.

It remains to show that

$$\int_{\mathbb{R}^4} b_0 \left(f_{A,B}(a, b) - f_{\tilde{A},\tilde{B}}(a, b)\right) da\, db \neq 0.$$

Step 3. As an intermediate step, we show the general formula

$$f_B(b) = \frac{-i}{(2\pi)^2}\int_{\mathbb{R}^2}\exp(-iv_0b_0)\frac{\partial\psi_{A,C}}{\partial u_1}(-v_0b_1,-v_1b_1,v_0,v_1)\,dv. \tag{31}$$

for the joint density f_B of B, if

$$\int_{\mathbb{R}^4} 1_{(-\infty,0]}(a_1)f_{A,C}(a,c)\,da\,dc = 0, \tag{32}$$

and if

$$\int_{\mathbb{R}^2}\sup_{u\in\mathbb{R}^2}\left|\partial_{u_1}\psi_{A,C}(u,v)\right|dv < \infty. \tag{33}$$

To show (31), choose a bivariate kernel function K which is absolutely-integrable, bounded by 1, satisfies $K(0)=1$ (e.g. an appropriately scaled normal density) and has an absolutely-integrable Fourier transform. Then from (4), using (32) we get that

$$f_B(b) = \int f_{A,B}(a,b)da = \int f_{A,C}\big(\tau(a,b)\big)|a_1|\,da = \lim_{\delta\downarrow 0}\int K(a\delta)f_{A,C}\big(\tau(a,b)\big)a_1\,da. \tag{34}$$

For any $\delta>0$ we compute, by Fourier inversion and Fubini's theorem, that

$$\begin{aligned}
&\int K(a\delta)f_{A,C}\big(\tau(a,b)\big)a_1\,da\\
&= (2\pi)^{-4}\int K(a\delta)\iint\exp\big(-i\tau(a,b)'(u,v)\big)\psi_{A,C}(u,v)a_1\,du\,dv\,da\\
&= (2\pi)^{-4}\int K(a\delta)\int\exp(-iv_0b_0)\int\exp\big(-ia_0(u_0+v_0b_1)\big)\\
&\qquad\qquad\cdot a_1\exp\big(-ia_1(u_1+v_1b_1)\big)\psi_{A,C}(u,v)\,du\,dv\,da\\
&= i(2\pi)^{-4}\iint\exp(-iv_0b_0)\Big(\int K(a\delta)\exp\big(-ia_0u_0\big)\big(\partial_{u_1}\exp\big(-ia_1u_1\big)\big)da\Big)\\
&\qquad\qquad\cdot\psi_{A,C}(u_0-v_0b_1,u_1-v_1b_1,v)\,du\,dv\\
&= -i(2\pi)^{-4}\iint\exp(-iv_0b_0)\delta^{-3}\big(\partial_{u_1}\mathscr{F}_2K\big)(-u/\delta)\,\psi_{A,C}(u_0-v_0b_1,u_1-v_1b_1,v)\,du\,dv\\
&= -i(2\pi)^{-4}\int\exp(-iv_0b_0)\int\big(\mathscr{F}_2K\big)(-u)\big(\partial_{u_1}\psi_{A,C}\big)(\delta u_0-v_0b_1,\delta u_1-v_1b_1,v)\,du\,dv,
\end{aligned}$$

by integration by parts in the last step, where we used

$$\begin{aligned}
&\int K(a\delta)\exp\big(-ia_0u_0\big)\big(\partial_{u_1}\exp\big(-ia_1u_1\big)\big)da\\
=&\,\partial_{u_1}\Big(\int K(a\delta)\exp\big(-ia_0u_0\big)\exp\big(-ia_1u_1\big)da\Big)\\
=&\,\partial_{u_1}\Big(\big(\mathscr{F}_2K(\cdot\delta)\big)(-u)\Big) = \partial_{u_1}\big((\mathscr{F}_2K)(-u/\delta)\big)/\delta^2\\
=&-\big(\partial_{u_1}(\mathscr{F}_2K)\big)(-u/\delta)/\delta^3.
\end{aligned}$$

Plugging this into (34) and letting $\delta\to 0$ and using dominated convergence, which is justified by (33), gives (31).

Step 4. We now apply (31) to (A,C) and $(\tilde{A},\tilde{C})$ having densities f_1 and f_2, and where the characteristic functions are determined by (30) and (29). We have already checked (32) for f_j. As for (33), consider for example the term $|u_0v_1\partial_{u_1}(\mathscr{F}_4G_1)(u,v)|$. Let $h=\mathscr{F}_1g_1$, an integrable function; then, since G_1 is a product density, we have

$$|u_0v_1\partial_{u_1}(\mathscr{F}_4G_1)(u,v)| = |u_0h(u_0)|\,|h'(u_1)||h(v_0)v_1\,h(v_1)|.$$

To bound $h'(u_1)$, relate this to the Fourier transform of the absolutely-integrable function $s\mapsto sg_1(s)$ so that h' is bounded. To bound $u_0h(u_0)$, relate this to the Fourier transform of g_1', which is also integrable, so that $u_0h(u_0)$ is bounded, and also integrable (and thus also $v_1\,h(v_1)$), as desired.

Next,

$$\psi_{\tilde{A},\tilde{C}}(u,v) - \psi_{A,C}(u,v) = 2\gamma \exp(3iu_1)\, Q\big(-i(u,v)\big)\, \big(\mathscr{F}_4 G_1\big)(u,v).$$

Since $\partial_{u_1} Q\big(-i(u,v)\big) = iv_0$, taking the derivative w.r.t. u_1 gives

$$\begin{aligned}\partial_{u_1}\big(\psi_{\tilde{A},\tilde{C}}(u,v) - \psi_{A,C}(u,v)\big) =& 2\gamma Q\big(-i(u,v)\big)\, \big(3i\exp(3iu_1)\big(\mathscr{F}_4 G_1\big)(u,v) + \exp(3iu_1)\,\partial_{u_1}\big(\mathscr{F}_4 G_1\big)(u,v)\big) \\ &+ 2iv_0\gamma \exp(3iu_1)\,\big(\mathscr{F}_4 G_1\big)(u,v), \qquad (u,v)\in\mathbb{R}^4.\end{aligned}$$

Since

$$Q\big(-i(-v_0b_1, -v_1b_1, v_0, v_1)\big) = 0,$$

applying (31) yields

$$f_B(b_0,b_1) - f_{\tilde{B}}(b_0,b_1) = \frac{-2i\gamma}{(2\pi)^2}\int_{\mathbb{R}^2} v_0 \exp(-iv_0b_0)\exp(-3iv_1b_1)\,\big(\mathscr{F}_4 G_1\big)(-v_0b_1, -v_1b_1, v)\,dv. \tag{35}$$

Step 5. In the final step, we show that

$$\int_{\mathbb{R}} b_0 \int_{\mathbb{R}} \big(f_B(b_0,b_1) - f_{\tilde{B}}(b_0,b_1)\big)\,db_1 db_0 \neq 0.$$

As above set $h = \mathscr{F}_1 g_1$, then since G_1 is a product density and g_1 and hence h are symmetric, we can rewrite the integral in (35) as the product

$$D(b_0,b_1) = \int_{\mathbb{R}} \exp(-iv_0b_0)h(v_0)\,v_0\,h(v_0b_1)\,dv_0 \int_{\mathbb{R}} \exp(-3iv_1b_1)\,h(v_1b_1)h(v_1)\,dv_1\,.$$

Using the Plancherel isometry, we evaluate the integral on the right as

$$\begin{aligned}E(b_1) &= \int_{\mathbb{R}} \exp(-3iv_1b_1)\,h(v_1)\,h(v_1b_1)\,dv_1 \\ &= 2\pi \int_{\mathbb{R}} \Big(\mathscr{F}_1^{-1}\big(\exp(-3ib_1\cdot)\,h(\cdot)\big)\Big)(t)\,\Big(\mathscr{F}_1^{-1}\big(h(\cdot b_1)\big)\Big)(t)\,dt \\ &= 2\pi \int_{\mathbb{R}} g_1(t+3b_1)g_1(t/b_1)/|b_1|\,dt \\ &= 2\pi \int_{\mathbb{R}} g_1(ub_1+3b_1)g_1(u)\,du.\end{aligned}$$

Let us discuss the function $E(b_1)$. Since g_1 is a bounded density, E is bounded by the maximal value of g_1 times 2π. Further, since g_1 has support $[-1,1]$, $g_1(ub_1+3b_1)$ has support $-3+[-1/b_1, 1/b_1]$. Therefore, E has compact support contained in $[-1/2, 1/2]$, and in particular is integrable. Further, $E \geq 0$ and in a neighborhood of zero, $E > 0$.

Now, since

$$\big|\exp(-iv_0b_0)h(v_0)\,v_0\,h(v_0b_1)\,E(b_1)\big| \leq \big|h(v_0)\,v_0\,E(b_1)\big|,$$

which is integrable, we may change the order of integration to obtain

$$\begin{aligned}F(b_0) := \int_{\mathbb{R}} D(b_0,b_1)\,db_1 &= \int_{\mathbb{R}} \exp(-iv_0b_0)h(v_0)\,v_0\,\Big(\int_{\mathbb{R}} h(v_0b_1)\,E(b_1)\,db_1\Big)\,dv_0 \\ &= 2\pi\,\big(\mathscr{F}^{-1}\tilde{F}\big)(b_0),\end{aligned}$$

where

$$\tilde{F}(v_0) = h(v_0)\,v_0 \int_{\mathbb{R}} h(v_0b_1)\,E(b_1)\,db_1.$$

We obtain

$$\int_{\mathbb{R}} b_0 \Big(\int_{\mathbb{R}} D(b_0,b_1)\,db_1 \Big) db_0 = \int_{\mathbb{R}} b_0 F(b_0)\,db_0 = (-i)\,\frac{\mathrm{d}}{\mathrm{d}v_0}\big(\mathscr{F}F\big)(0)$$
$$= 2\pi(-i)\,\frac{\mathrm{d}}{\mathrm{d}v_0}\tilde{F}(0) = 2\pi(-i)\int_{\mathbb{R}} E(b_1)\,db_1 \neq 0,$$

since $h(0)=1$ and h' is bounded, which concludes the proof. ∎

Theorem 8. *Consider the triangular model (1) under Assumption 1, and suppose that $L=S=1$. Then, the mean of B_2 can not be identified from the distribution of the observations (X,Y,Z,W), even if (Z,W) has full support, if all infinitely differentiable densities with compact support are admitted as the joint density of $(A_0,A_1,A_2,B_0,B_1,B_2)'$.*

Proof of Theorem 8. By replacing A_0 by A_2 and C_0 by C_2, from the counterexample of Theorem 1, there exist two distinct infinitely differentiable densities f_j, $j=1,2$ with compact support in $[1,\infty)\times\mathbb{R}^3$, for which $\mathscr{F}_4(f_2-f_1)$ vanishes on the set $\{(u_1,u_2,v_1,v_2)'\in\mathbb{R}^4 : \quad Q(-i(u_1,u_2,v_1,v_2)')=0\}$, and for which

$$\int_{\mathbb{R}} b_2 \int_{\mathbb{R}^2}\int_1^\infty a_1\big(f_1-f_2\big)(a_1,a_2,a_1b_1,b_2+b_1a_2)\,da_1da_2db_1db_2 \neq 0. \tag{36}$$

Now, consider the two densities

$$f_{A,C;j}(a,c) = g_1(a_0)\,g_1(c_0)\,f_j(a_1,a_2,c_1,c_2)$$

for (A,C), $A=(A_0,A_1,A_2)'$ and $C=(C_0,C_1,C_2)'$, and where g_1 is defined in (28). The corresponding densities of (A,B) are given by

$$f_{A,B;j}(a,b) = a_1\,g_1(a_0)\,g_1(b_0+b_1a_0)\,f_j(a_1,a_2,b_1a_1,b_2+b_1a_2),$$

with marginal densities of B_2

$$\begin{aligned} f_{B_2;j}(b_2) &= \int_{\mathbb{R}^5} f_{A,B;j}(a,b)\,da\,db_0\,db_1 \\ &= \int_{\mathbb{R}^5} a_1\,f_j(a_1,a_2,b_1a_1,b_2+b_1a_2)\Big[\int_{\mathbb{R}} g_1(a_0)\Big(\int_{\mathbb{R}} g_1(b_0+b_1a_0)\,db_0\Big)da_0\Big]da_1\,da_2\,db_1 \\ &= \int_{\mathbb{R}^3} a_1\,f_j(a_1,a_2,a_1b_1,b_2+b_1a_2)\,da_1\,da_2\,db_1 \end{aligned}$$

so that

$$\int_{\mathbb{R}} b_2\big(f_{B_2;1}-f_{B_2;2}\big)(b_2)\,db_2 \neq 0$$

by (36). It remains to show that both densities lead to the same distribution of the observed random variables (Y,X,Z,W). Since (Z,W) are exogenous, as in Lemma 2.1 it suffices to show that the conditional characteristic function of (Y,X) given $W=w,Z=z$ coincide for all $w,z\in\mathbb{R}$. Indeed,

$$E\Big(\exp\big(it_1X+it_2Y\big)\Big|Z=z,W=w\Big) = \psi_{A,C}(t_1,t_1z,t_1w,t_2,t_2z,t_2w),$$

and, setting $h=\mathscr{F}_1g_1$,

$$\big(\psi_{A,C;2}-\psi_{A,C;1}\big)(t_1,t_1z,t_1w,t_2,t_2z,t_2w) = h(t_1)h(t_2)\big(\mathscr{F}_4(f_2-f_1)\big)(t_1z,t_1w,t_2z,t_2w) = 0$$

since $Q\big(-i(t_1z,t_1w,t_2z,t_2w)\big)=0$. ∎

Proof of Proposition 2.1. Let $V=\big(EY,EYZ,EYW\big)'$ and

$$M = E\big(1,Z,W\big)'\big(1,X,W\big) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & EZX & EZW \\ 0 & EWX & EW^2 \end{pmatrix},$$

so that $\mu_{IV} = M^{-1}V$. Using exogeneity and normalization, we immediately compute

$$V = \begin{pmatrix} EB_0 + EB_1X \\ EB_1XZ + EB_2EZW \\ EB_1XW + EB_2EW^2 \end{pmatrix} \tag{37}$$

Further,

$$\begin{aligned} EB_1X &= EB_1A_0, \\ EB_1XZ &= EA_1B_1EZ^2 + EB_1A_2EZW, \\ EB_1XW &= EA_1B_1EZW + EB_1A_2EW^2. \end{aligned} \tag{38}$$

Similarly, for the entries of the matrix M we compute

$$\begin{aligned} EXZ &= EA_1EZ^2 + EA_2EZW, \\ EXW &= EA_1EZW. \end{aligned} \tag{39}$$

Inserting (39) into M we compute

$$\det M = EA_1\left(EZ^2EW^2 - \left(EZW\right)^2\right),$$

so that M is invertible and a straightforward computation using (37) - (39) leads to the formula for μ_{IV}. ∎

A.2. Proofs for Section 3

Proof of (7). Observe that

$$\begin{aligned} f_A(a_0,a_1) &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}\exp\left(-i(a_0u_0+a_1u_1)\right)\left(\mathscr{F}_2f_A\right)(u_0,u_1)du_0\,du_1 \\ &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}|t|\exp\left(-it\,(a_0+a_1z)\right)\left(\mathscr{F}_2f_A\right)(t,tz)dt\,dz \\ &= \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2}|t|\exp\left(-it\,(a_0+a_1z)\right)\mathscr{F}_1\left(f_{X|Z}\right)(t,z)\,dt\,dz, \end{aligned}$$

that is, (7). ∎

The following smoothness assumption is needed in the second part of Theorem 3 to justify interchanging the orders of integration.

Assumption 10.

$$\int_{\mathbb{R}^3}|t|\left|\int_{\mathbb{R}^3}\exp\left(it(b_0+b_1x)\right)f_{B,A_0,A_1}(b,x-a_1z,a_1)\,da_1\,db_0\,db_1\right|dz\,dt\,dx<\infty.$$

Proof of Theorem 3. (i) We provide the missing details for the proof of the first part of Theorem 3. By simple change of variables, for fixed z,

$$f_{B,A_0+A_1z,A_1}(b,x,a_1) = f_{B,A_0,A_1}(b,x-a_1z,a_1),$$

so that

$$f_{B,A_0+A_1z}(b,x) = \int_{\mathbb{R}}f_{B,A_0,A_1}(b,x-a_1z,a_1)\,da_1$$

and therefore

$$\begin{aligned} &E\left(\exp\left(it(B_0+B_1x)|A_0+A_1z=x\right)\right) f_{A_0+A_1z}(x) \\ &=\int_{\mathbb{R}^3}\exp\left(it(b_0+b_1x)\right) f_{B,A_0,A_1}(b,x-a_1z,a_1)\,da_1\,db_0\,db_1. \end{aligned}$$

Moreover, by exogeneity of Z,

$$f_{A_0+A_1z}(x)=f_{X|Z}(x|z),$$

so that by (9) and the above two displays,

$$\int_{\mathbb{R}}\mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z)\,f_{X|Z}(x|z)\,dz=\int_{\mathbb{R}^4}\exp\left(it(b_0+b_1x)\right) f_{B,A_0,A_1}(b,x-a_1z,a_1)\,dz\,da_1\,db_0\,db_1. \tag{40}$$

The computation in (10) and Assumption 3 then lead to (11). As in the above proof of (7), applying the operator T now yields the density f_B, and we get the reconstruction formula (12).

The expression (13) for $E|A_1|^{-1}$ is obtained by setting $t=0$ in (11).

(ii) By (40), Assumption 10 is equivalent to

$$\int_{\mathbb{R}^3}|t|\left|\mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z)\right| f_{X|Z}(x|z)\,dz\,dt\,dx<\infty.$$

By inserting the definition of T and changing the order of integration, we obtain

$$\begin{aligned} &T\Big(\int_{\mathbb{R}}\mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z)\,f_{X|Z}(x|z)\,dz\Big)(b_0,b_1) \\ &=\frac{1}{(2\pi)^2}\int_{\mathbb{R}}\int_{\mathbb{R}}|t|e^{-itb_0}\Big(\int_{\mathbb{R}}\mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z)\,e^{-itb_1x}f_{X|Z}(x|z)\,dx\Big)\,dt\,dz. \end{aligned}$$

Using

$$\psi_{X,Y|Z}(t_1,t_2|z)=\int_{\mathbb{R}}\mathscr{F}_1\left(f_{Y|X,Z}\right)(t_2|x,z)e^{it_1x}f_{X|Z}(x|z)\,dx.$$

yields (12). ∎

Proof of Theorem 4. First we observe that by Assumption 2, the support of X is also $\mathbb{R}$. Now, we start by showing that for the characteristic function of B,

$$\begin{aligned} &\left(\mathscr{F}_{2+S}f_B\right)(t,tx,tw)\,E\left|A_{1,1}\right|^{-1}=E\left(\exp\left(it(B_0+B_1x+B_2'w)\right)\right)E\left|A_{1,1}\right|^{-1} \\ &=\int_{\mathbb{R}^L}\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)\,f_{X|Z,W}(x|z,w)\,f_{Z_{-1}}(z_{-1})\,dz. \end{aligned} \tag{41}$$

To this end, using the exogeneity Assumption 2 we compute that

$$\begin{aligned} \mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)=&E\left(\exp\left(it(B_0+B_1X+B_2'W)\right)|X=x,Z=z,W=w\right) \\ =&E\left(\exp\left(it(B_0+B_1x+B_2'w)\right)|A_0+A_1'z+A_2'w=x\right). \end{aligned} \tag{42}$$

Since

$$\begin{aligned} f_{B,A_0+A_1'z+A_2'w,A_1,A_2}(b,x,a_1,a_2)&=f_{B,A_0,A_1,A_2}(b,x-a_1'z-a_2'w,a_1,a_2), \\ f_{A_0+A_1'z+A_2'w}(x)&=f_{X|Z,W}(x|z,w), \end{aligned}$$

we obtain that

$$\begin{aligned} &\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)\,f_{X|Z,W}(x|z,w) \\ &=\int_{\mathbb{R}^{2+2S+L}}\exp\left(it(b_0+b_1x+b_2'w)\right)f_{B,A}(b,x-a_1'z-a_2'w,a_1,a_2)\,da_1\,da_2\,db. \end{aligned} \tag{43}$$

Integrating out z_1 gives

$$\begin{aligned}
&\int_{\mathbb{R}} \mathscr{F}_1\big(f_{Y|X,Z,W}\big)(t|x,z,w)\, f_{X|Z,W}(x|z,w)\, dz_1 \\
&\quad = \int_{\mathbb{R}^{3+2S+L}} \exp\big(it(b_0+b_1x+b_2'w)\big)\, f_{B,A}(b,x-a_1'z-a_2'w,a_1,a_2)\, dz_1\, da_1\, da_2\, db \\
&\quad = \int_{\mathbb{R}^{3+2S+L}} |a_{1,1}|^{-1} \exp\big(it(b_0+b_1x+b_2'w)\big)\, f_{B,A}(b,a)\, da\, db \\
&\quad = E\Big(\exp\big(it(B_0+B_1x+B_2w)\big)\,\frac{1}{|A_{1,1}|}\Big) = E\Big(\exp\big(it(B_0+B_1x+B_2w)\big)\Big)\, E|A_{1,1}|^{-1}.
\end{aligned} \tag{44}$$

using a change of variables and Assumption 5 in the last step. Averaging over the values of Z_{-1} then gives (41), and applying the operator T_{S+1}, by proceeding as in the proof of (7) we obtain (14). Finally, taking $t=0$ in (44), we see that for any x, z_{-1}, w, (15) holds true. ∎

Proof of Proposition 3.1. (i). We have $1_{B_0\le t}\, C_{A_1}^{-1} \le 1_{B_0\le t}\,|A_1|^{-1} \le 1_{B_0\le t}\, c_{A_1}^{-1}$. Taking expected values, observing (16) and solving for $F_{B_0}(t)$ gives (i).
(ii). From the Hölder inequality, we obtain

$$F(t) = E\Big(1_{B_0\le t}\,|A_1|^{-1}\Big) \le \Big(E\big(|A_1|^{-p}\big)\Big)^{1/p}\big(F_{B_0}(t)\big)^{(p-1)/p}.$$

Solving for $F_{B_0}(t)$ gives (ii).
(iii). Set $r=(p+1)/p$, so that $p=1/(r-1)$. Apply the Hölder inequality with exponents r and $s=r/(r-1)$ to obtain

$$\begin{aligned}
F_{B_0}(t) &= E\Big(\big(1_{B_0\le t}\,|A_1|^{-1}\big)^{1/r}|A_1|^{1/r}\big)\Big) \\
&\le \big(F(t)\big)^{1/r} E\big(|A_1|^{s/r}\big)^{1/s} = \big(F(t)\big)^{p/(p+1)}\, E\big(|A_1|^{p}\big)^{1/(p+1)}.
\end{aligned}$$

∎

Proof of Theorem 5. For a given $x\in \operatorname{supp} X$ consider the random variable $A_{x,w} = (x-A_0-A_2'w)/A_1$. By a change of variables,

$$f_{A_x,A_1,A_2,B}(z,a_1,a_2,b) = |a_1|\, f_{A_0,A_1,A_2,B}(x-a_1z-a_2'w,a_1,a_2,b).$$

Therefore, from (43) we obtain that

$$\begin{aligned}
&\mathscr{F}_1\big(f_{Y|X,Z,W}\big)(t|x,z,w)\, f_{X|Z,W}(x|z,w) \\
&\quad = \int_{\mathbb{R}^{3+2S}} \exp\big(it(b_0+b_1x+b_2'w)\big)\,|a_1|^{-1}\, f_{A_x,A_1,A_2,B}(z,a_1,a_2,b)\, da_1\, da_2\, db.
\end{aligned}$$

Under the support assumption (17), it suffices to integrate out z over the support of the conditional distribution of $Z|W=w$ to obtain (18). ∎

Proof of Theorem 6. First, we require the following lemma, which does not involve the model itself but only a set of random coefficients.

Lemma A.1. *Let $(A',B')'$, $A'=(A_0,A_1)$, $B'=(B_0,B_1)$ be a four-dimensional random vector with continuous Lebesgue density, which satisfies Assumptions 3 and 10, and for which $\mathscr{F}_2 f_B$ is integrable. Set $C_0=B_0+B_1A_0$, $C_1=B_1A_1$ and $C'=(C_0,C_1)$, and let $\psi_{A,C}$ denote the characteristic function of $(A',C')'$. Then*

$$f_B(b_0,b_1) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}}\int_{\mathbb{R}} \exp(-itb_0)\,\psi_{A,C}(-tb_1,-tzb_1,t,tz)\,|t|\,dt\,dz\,\big(E|A_1|^{-1}\big)^{-1}. \tag{45}$$

Proof. Choose any Z with full support, and independent of $(A',B')'$, and form model (2) (that is, define Y,X according to (2)). Then the assumptions of Theorem 3, *(ii)*, are satisfied, and we obtain (12). Using the equality (5) immediately gives (45). ∎

The following lemma is based on analytic continuation.

Lemma A.2. *Under Assumption 7, for any fixed t and b_1 the function*

$$\Phi : z \mapsto \psi_{A,C}(-tb_1, -tb_1 z, t, tz),$$

is uniquely determined by its restriction to a non-empty interval.

Proof of Lemma A.2: Suppose that A, C and $\tilde{A}, \tilde{C}$ are two-dimensional random vectors both of which satisfy Assumption 7. Suppose that for fixed t and b_1 the functions

$$\Phi_0 : z \mapsto \psi_{A,C}(-tb_1, -tb_1 z, t, tz), \qquad \Phi_1 : z \mapsto \psi_{\tilde{A},\tilde{C}}(-tb_1, -tb_1 z, t, tz)$$

coincide on the non-void interval I. Let $\Phi := \Phi_0 - \Phi_1$ and $\psi = \psi_{A,C} - \psi_{\tilde{A},\tilde{C}}$, we need to show that Φ vanishes identically.

First we show that the function Φ can be represented by its Taylor series around the center z_0 of I. The residual term $R_k(z)$ of the kth Taylor polynomial of Φ obeys the bound

$$|R_k(z)| \leq \frac{1}{(k+1)!} |z - z_0|^{k+1} \left\| \Phi^{(k+1)} \right\|_\infty,$$

where we write $\Phi^{(k)}$ for the kth derivative of Φ. We deduce that

$$\Phi^{(k+1)}(z) = \sum_{l=0}^{k+1} \binom{k+1}{l} (-tb_1)^l t^{k+1-l} \frac{\partial^{k+1} \psi}{(\partial a_1)^l (\partial c_1)^{k+1-l}} (-tb_1, -tb_1 z, t, tz).$$

Since

$$\left| \frac{\partial^{k+1} \psi}{(\partial a_1)^l (\partial c_1)^{k+1-l}} (-tb_1, -tb_1 z, t, tz) \right| \leq E|A_1|^l |C_1|^{k+1-l} + E|\tilde{A}_1|^l |\tilde{C}_1|^{k+1-l},$$

it follows by binomial expansion that

$$\begin{aligned}
|\Phi^{(k+1)}(z)| &\leq \sum_{l=0}^{k+1} \binom{k+1}{l} |tb_1|^l |t|^{k+1-l} \left(E|A_1|^l |C_1|^{k+1-l} + + E|\tilde{A}_1|^l |\tilde{C}_1|^{k+1-l} \right) \\
&= |t|^{k+1} \left(E\left(|b_1 A_1| + |C_1| \right)^{k+1} + E\left(|b_1 \tilde{A}_1| + |\tilde{C}_1| \right)^{k+1} \right) \\
&\leq 2^{k+1} |tb_1|^{k+1} \left(E|A_1|^{k+1} + E|\tilde{A}_1|^{k+1} \right) + 2^{k+1} |t|^{k+1} \left(E|C_1|^{k+1} + E|\tilde{C}_1|^{k+1} \right).
\end{aligned}$$

By Assumption 7 we conclude that

$$\lim_{k \to \infty} R_k(z) = 0,$$

for all $z \in \mathbb{R}$, which yields pointwise convergence of the Taylor series to Φ on the whole real line.

The function Φ vanishes on I, thus on some non-void open interval around z_0. Therefore all derivatives of Φ at z_0 equal zero so that the Taylor series of Φ around z_0 converges to zero everywhere. Combining this with the above paragraph we conclude that $\Phi \equiv 0$ and, hence, $\Phi_0 = \Phi_1$ throughout. This completes the proof of the lemma. □

Proof of Theorem 6 continued.

From (5), for any fixed t and b_1 we identify the function

$$\Phi : z \mapsto \psi_{A,C}(-tb_1, -tb_1 z, t, tz) = \psi_{X,Y|Z}(-tb_1, t; z)$$

over the support $\mathscr{S}_Z$. From Lemma A.2, we hence identify $\psi_{A,C}(-tb_1, -tb_1 z, t, tz)$ for all z, and therefore, we identify the function

$$g(b_0, b_1) = \frac{1}{(2\pi)^2} \iint \exp(-itb_0) \psi_{A,C}(-tb_1, -tzb_1, t, tz) \, |t| \, dt \, dz.$$

Since by Corollary 45,

$$f_B = g/\int g,$$

we identify f_B. ∎

B. Semiparametric estimation: Technical assumptions, results and proofs

Lemma B.1. *Let $\Gamma(x)$ denote the gamma function. For any $\kappa \geq 2$ and $n \geq 2$, we have under Assumption 8 that*

$$E\sum_{j=1}^{n-1}\left(Z_{(j+1)} - Z_{(j)}\right)^{\kappa} \leq n(n-1)^{-\kappa} c_Z^{-\kappa}\,\kappa\Gamma(\kappa),$$
$$\max\left\{E(1-Z_{(n)})^{\kappa}, E(Z_{(1)}+1)^{\kappa}\right\} \leq \kappa c_Z^{-\kappa} n^{-\kappa}\Gamma(\kappa).$$

The proof is defered to the technical supplement.

Assumption 11. There exists a $C_A > 0$ such that

$$\sup_{x\in I, z\in[-1,1]}\int_{\mathbb{R}} f_{A_0,A_1}\left(x - a_1 z, a_1\right) da_1 \leq C_A,$$

as well as

$$\sup_{x\in I}\left|\int_{\mathbb{R}}\left[f_{A_0,A_1}\left(x-a_1 z, a_1\right) - f_{A_0,A_1}\left(x - a_1 w, a_1\right)\right] da_1\right| \leq C_A|z-w| \quad z,w\in[-1,1].$$

Assumption 12.

$$\sup_{x\in I}\int_{\mathbb{R}^3}\left|f_{A_0,A_1,B}\left(x-a_1 z, a_1, b\right) - f_{A_0,A_1,B}\left(x-a_1 w, a_1, b\right)\right] da_1\, db\Big| \leq C_A|z-w| \quad z,w\in[-1,1].$$

Proof of Proposition 4.1. For brevity, in this proof let

$$s_c := E|A_1|^{-1} = \int_I v(x)\int_{-1}^{1} f_{X|Z}(x|z)\,dz\,dx,$$

Let σ_Z denote the σ-algebra generated by the Z_i, $i = 1,\ldots,n$. Then

$$E\left(\hat{a}_{I,n} - s_c\right)^2 = E\,\mathrm{Var}\left(\hat{a}_{I,n}|\sigma_Z\right) + E\left[E\left(\hat{a}_{I,n}|\sigma_Z\right) - s_c\right]^2. \tag{46}$$

Since the $X_{(j)}$ are independent conditional on σ_Z, we have

$$\begin{aligned}\mathrm{Var}\left(\hat{a}_{I,n}|\sigma_Z\right) &\leq \sum_{j=1}^{n-1} E\left(v^2(X_{(j)})\right)|\sigma_Z\Big)\left(Z_{(j+1)} - Z_{(j)}\right)^2 \\ &\leq C_v^2\sum_{j=1}^{n-1}\left(Z_{(j+1)} - Z_{(j)}\right)^2,\end{aligned}$$

where C_v is a bound for v. From Lemma B.1,

$$E\mathrm{Var}\left(\hat{a}_{I,n}|\sigma_Z\right) \leq 2C_v^2 c_Z^{-2} n(n-1)^{-2}. \tag{47}$$

Further, we have that

$$\begin{aligned}E\left(\hat{a}_{I,n}|\sigma_Z\right) &= \int_I\int_{-1}^{1} v(x)\,\tilde{f}(x,z)\,dz\,dx, \\ \tilde{f}(x,z) &= \sum_{j=1}^{n-1} f_{X|Z}\left(x|Z_{(j)}\right)1_{[Z_{(j)},Z_{(j+1)})}(z).\end{aligned}$$

Using the Cauchy-Schwarz inequality twice, we estimate

$$\begin{aligned}
&E\left[E\left(\hat{a}_{I,n}|\sigma_Z\right)-s_c\right]^2\\
\leq& 2E\sum_{j=1}^{n-1}\int_I\int_{Z_{(j)}}^{Z_{(j+1)}}\left[f_{X|Z}\left(x|z\right)-f_{X|Z}\left(x|Z_{(j)}\right)\right]^2 dz\,v(x)\,dx\\
&+2E\left(\int_I\int_{-1}^{Z_{(1)}} f^2_{X|Z}\left(x|z\right)dz\,v(x)\,dx+\int_I\int_{Z_{(n)}}^{1} f^2_{X|Z}\left(x|z\right)dz\,v(x)\,dx\right).
\end{aligned}$$

For the second term, using Assumption 11 and Lemma B.1, we estimate

$$\begin{aligned}
&E\int_I\int_{-1}^{Z_{(1)}} f^2_{X|Z}\left(x|z\right)dz\,v(x)\,dx+E\int_I\int_{Z_{(n)}}^{1} f^2_{X|Z}\left(x|z\right)dz\,v(x)\,dx\\
\leq& C_A^2\,E\left(Z_{(1)}+1+1-Z_{(n)}\right)\leq 2C_A^2\,c_Z^{-1}n^{-1}.
\end{aligned}$$

For the first term, we observe that for $x\in[Z_{(j)},Z_{(j+1)})$, again by Assumption 11,

$$\begin{aligned}
\left|f_{X|Z}\left(x|z\right)-f_{X|Z}\left(x|Z_{(j)}\right)\right| &\leq \left|\int_{\mathbb{R}}\left(f_{A_0,A_1}(x-za_1,a_1)-f_{A_0,A_1}(x-Z_{(j)}a_1,a_1)\right)da_1\right|\\
&\leq C_A\left|z-Z_{(j)}\right|
\end{aligned}$$

so that using Lemma B.1,

$$\begin{aligned}
&E\sum_{j=1}^{n-1}\int_I v(x)\int_{Z_{(j)}}^{Z_{(j+1)}}\left[f_{X|Z}\left(x|z\right)-f_{X|Z}\left(x|Z_{(j)}\right)\right]^2 dz\,dx\\
\leq& E\sum_{j=1}^{n-1}\left(Z_{(j+1)}-Z_{(j)}\right)^3 C_A^2/3\leq 2C_A^2\,c_Z^{-3}\,n\,(n-1)^{-3}.
\end{aligned}$$

Together with (47) and (46) this gives the statement of the proposition. ∎

Proof of Theorem 7. By θ_n we denote an element of Θ_n closest to the true θ_0 (with respect to the Euclidean distance). The definition and the geometric structure of Θ_n yield that

$$\|\theta_0-\theta_n\|^2\leq d\,b^2\,n^{-1}\,. \tag{48}$$

Given $\varepsilon>0$, we need to find $M>0$ such that

$$P\left[\|\hat{\theta}_n-\theta_0\|>M/\sqrt{n}\right]\leq\varepsilon,$$

for all (sufficiently large) $n\in\mathbb{N}$. We choose $\tilde{M}>0$ so large that

$$\begin{aligned}
&P\left[\|\hat{\theta}_n-\theta_0\|>M/\sqrt{n}\right]\\
\leq& P\left[\|\hat{\theta}_n-\theta_0\|>M/\sqrt{n},\ \left|\hat{a}_{I;n}-E|A_1|^{-1}\right|\leq\tilde{M}/\sqrt{n}\right]+P\left[\left|\hat{a}_{I;n}-E|A_1|^{-1}\right|>\tilde{M}/\sqrt{n}\right]\\
\leq& P\left[\|\hat{\theta}_n-\theta_0\|>M/\sqrt{n},\ \left|\hat{a}_{I;n}-E|A_1|^{-1}\right|\leq\tilde{M}/\sqrt{n}\right]+\varepsilon/2
\end{aligned}$$

for all n. Let

$$C_\Theta=\sup_{\theta\in\Theta}\|\Phi(\theta,\cdot)\|_{v;q},$$

for which $0 < C_\Theta < \infty$, and for the rest of the proof let $s_c = E|A_1|^{-1}$. Then

$$\begin{aligned}
&P\big[\|\hat\theta_n - \theta_0\| > M/\sqrt{n},\ \big|\hat a_{I;n} - s_c\big| \le \tilde M/\sqrt{n}\big] \\
\le& P\Big[\exists \theta' \in \Theta_n : \|\theta' - \theta_0\| > M/\sqrt{n}, \|\hat\Phi_n(\cdot) - \hat a_{I;n}\Phi(\theta',\cdot)\|_{v;q} \le \|\hat\Phi_n(\cdot) - \hat a_{I;n}\Phi(\theta_n,\cdot)\|_{v;q}, \\
&\qquad \big|\hat a_{I;n} - s_c\big| \le \tilde M/\sqrt{n}\Big] \\
\le& P\Big[\exists \theta' \in \Theta_n : \|\theta' - \theta_0\| > M/\sqrt{n}, \quad \|\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} \\
&\qquad\qquad \le \|\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} + 2C_\Theta\tilde M/\sqrt{n}\Big] \\
\le& \sum_{\theta'\in\Theta_n} 1_{(M/\sqrt{n},\infty)}(\|\theta' - \theta_0\|) P\Big[\|\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} \le \|\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} + 2C_\Theta\tilde M/\sqrt{n}\Big]
\end{aligned}$$

by the triangle inequality. Now,

$$\|\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} \le \|\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} + 2C_\Theta\tilde M/\sqrt{n}$$

implies

$$\begin{aligned}
&\Big|\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q}\Big| \\
\le& \|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q} + \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} + 2C_\Theta\tilde M/\sqrt{n}
\end{aligned}$$

and hence

$$\begin{aligned}
&\max\Big(\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} - 2C_\Theta\tilde M/\sqrt{n}, 0\Big) \\
\le& 2\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q}.
\end{aligned}$$

Therefore

$$\begin{aligned}
&P\big[\|\hat\theta_n - \theta_0\| > M/\sqrt{n},\ \big|\hat a_{I;n} - s_c\big| \le \tilde M/\sqrt{n}\big] \\
\le& \sum_{\theta'\in\Theta_n} 1_{(M/\sqrt{n},\infty)}(\|\theta' - \theta_0\|) P\Big[\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q} \ge \\
&\qquad \frac{1}{2}\max\Big(\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} - 2C_\Theta\tilde M/\sqrt{n}, 0\Big)\Big] \\
\le& 2^\kappa \sum_{\theta'\in\Theta_n} 1_{(M/\sqrt{n},\infty)}(\|\theta' - \theta_0\|) E\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|^\kappa_{v;q} \\
&\qquad \max\Big(\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} - 2C_\Theta\tilde M/\sqrt{n}, 0\Big)^{-\kappa}
\end{aligned} \tag{49}$$

for some fixed integer $\kappa \ge 2$ (to be specified later) by the Markov inequality. Below we show the estimates

$$E\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|^\kappa_{v;q} = \mathscr{O}(n^{-\kappa/2}), \tag{50}$$

$$\begin{aligned}
&\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} - 2C_\Theta\tilde M/\sqrt{n} \\
\ge& c^{1/2}_{\Theta,0}\|\theta_0 - \theta'\| - \big(c^{1/2}_{\Theta,1}\sqrt{d}b + 2C_\Theta\tilde M\big)n^{-1/2} - \mathscr{O}(1/n),
\end{aligned} \tag{51}$$

If we set

$$\alpha = \big[c^{1/2}_{\Theta,0}M - \big(c^{1/2}_{\Theta,1}\sqrt{d}b + 2C_\Theta\tilde M + 1\big)\big],$$

then $\alpha \to \infty$ as $M \to \infty$, and for large n and $\|\theta - \theta'\| > M/\sqrt{n}$, from (51)

$$\Big|\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta',\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_n,\cdot)\|_{v;q} - 2C_\Theta\tilde M/\sqrt{n}\Big| \ge \alpha/\sqrt{n}.$$

On the other hand, (51) is also bounded from below by

$$c_{\Theta,0}^{1/2}\|\theta_n-\theta'\|-\big(c_{\Theta,1}^{1/2}+c_{\Theta,0}^{1/2}\big)\sqrt{d}b+2C_\Theta\tilde{M}\big)n^{-1/2}-\mathscr{O}(1/n)$$
$$\geq c_{\Theta,0}^{1/2}bn^{-1/2}J(\theta')-\big(c_{\Theta,1}^{1/2}+c_{\Theta,0}^{1/2}\big)\sqrt{d}b+2C_\Theta\tilde{M}\big)n^{-1/2}-\mathscr{O}(1/n),$$

for some integer $J(\theta')\geq 0$ where the preimages $J^{-1}(\{j\})$ contain at most $2(2j+1)^{d-1}$ elements for all integer $j\geq 0$. Therefore, from (49), (50) and the above estimates,

$$\begin{aligned}
&P\big[\|\hat{\theta}_n-\theta_0\|>M/\sqrt{n},\ \big|\hat{a}_{I;n}-s_c\big|\leq\tilde{M}/\sqrt{n}\big]\\
&\leq\sum_{j\geq 0}(2j+1)^{d-1}\cdot\mathscr{O}\big(n^{-\kappa/2}\big)\\
&\qquad\big(\max\big\{\alpha n^{-1/2},c_{\Theta,0}^{1/2}bn^{-1/2}j-\big(c_{\Theta,1}^{1/2}+c_{\Theta,0}^{1/2}\big)\sqrt{d}b+2C_\Theta\tilde{M}\big)n^{-1/2}-\mathscr{O}(1/n)\big\}\big)^{-\kappa}\\
&=\mathscr{O}(1)\cdot\sum_{j\geq 0}\big(\max\big\{\alpha,c_{\Theta,0}^{1/2}bj-\big(c_{\Theta,1}^{1/2}+c_{\Theta,0}^{1/2}\big)\sqrt{d}b+2C_\Theta\tilde{M}\big)-\mathscr{O}(n^{-1/2})\big\}\big)^{-\kappa}\cdot(2j+1)^{d-1},
\end{aligned}$$

where the constants contained in $\mathscr{O}(\cdot)$ depend on neither α nor n. We choose $\kappa>d$. Then, by the dominated convergence theorem, we deduce that the above expression tends to zero as $\alpha\uparrow\infty$ (or $M\uparrow\infty$) – uniformly with respect to n.

It remains to prove (50) and (51). Using the simple inequality $(a+b)^\kappa\leq 2^\kappa(a^\kappa+b^\kappa)$, $a,b>0$, $\kappa\in\mathbb{N}$, we estimate

$$\begin{aligned}
&E\big\|\hat{\Phi}_n(\cdot)-E\hat{\Phi}_n(\cdot)\big\|_{\nu;q}^\kappa\leq\frac{1}{q}\sum_{p=1}^q\int_{\mathbb{R}}\big|\hat{\Phi}_n(t,I_p)-E\hat{\Phi}_n(t,I_p)\big|^\kappa d\nu(t)\\
&\leq\frac{2^\kappa}{q}\sum_{p=1}^q\int_{\mathbb{R}}\Big(\big|\hat{\Phi}_n(t,I_p)-E\big(\hat{\Phi}_n(t,I_p)|\sigma_Z\big)\big|^\kappa+\big|\mathbb{E}\big(\hat{\Phi}_n(t,I_p)|\sigma_Z\big)-E\hat{\Phi}_n(t,I_p)\big|^\kappa\Big)d\nu(t),
\end{aligned}\tag{52}$$

We have that

$$\begin{aligned}
E\big(\hat{\Phi}_n(t,I_p)\mid\sigma_Z\big)&=\sum_{j=1}^{n-1}E\Big(\exp(itY_{(j)})1_{I_p}(X_{(j)})\big|\sigma_Z\Big)\cdot\big(Z_{(j+1)}-Z_{(j)}\big)\\
&=\sum_{j=1}^{n-1}\big(Z_{(j+1)}-Z_{(j)}\big)\iint\exp(ity')1_{I_p}(x')\,f_{Y,X|Z}(y',x'|Z_{(j)})\,dy'\,dx'\\
&=\sum_{j=1}^{n-1}\int_{z=Z_{(j)}}^{Z_{(j+1)}}\iint\exp(ity')1_{I_p}(x')\,f_{Y,X|Z}(y',x'|Z_{(j)})\,dy'\,dx'\,dz.
\end{aligned}\tag{53}$$

For all $t\in\mathbb{R}$, $s>0$, from the Hoeffding inequality,

$$P\big[\big|\hat{\Phi}_n(t,I_p)-E(\hat{\Phi}_n(t,I_p)\mid\sigma_Z)\big|>s\mid\sigma_Z\big]\leq 4\exp\Big\{-\frac{1}{8}s^2/\Big(\sum_{j=1}^{n-1}(Z_{(j+1)}-Z_{(j)})^2\Big)\Big\},$$

Therefore

$$\begin{aligned}
E\big(\big|\hat{\Phi}_n(t,I_p)-E(\hat{\Phi}_n(t,I_p)\mid\sigma_Z)\big|^\kappa\mid\sigma_Z\big)&\leq 4\int_{s>0}\exp\Big\{-\frac{1}{8}s^{2/\kappa}/\Big(\sum_{j=1}^{n-1}(Z_{(j+1)}-Z_{(j)})^2\Big)\Big\}ds\\
&\leq 2\kappa\sqrt{8}^\kappa\Gamma(\kappa/2)\cdot\Big(\sum_{j=1}^{n-1}(Z_{(j+1)}-Z_{(j)})^2\Big)^{\kappa/2}\\
&\leq 2\kappa\sqrt{8}^\kappa\Gamma(\kappa/2)\cdot(n-1)^{\kappa/2-1}\cdot\sum_{j=1}^{n-1}(Z_{(j+1)}-Z_{(j)})^\kappa,
\end{aligned}$$

again by Hölder's inequality. Taking the expectation on both sides of the above inequality and using Lemma B.1

yields uniformly in t that

$$E\big|\hat{\Phi}_n(t,I_p) - E(\hat{\Phi}_n(t,I_p) \mid \sigma_Z)\big|^\kappa = \mathcal{O}(n^{-\kappa/2}). \tag{54}$$

To proceed, first observe that

$$\big|E\hat{\Phi}_n(t,I_p) - s_c\Phi(\theta_0,t,I_p)\big|^\kappa \le E\big|E(\hat{\Phi}_n(t,I_p) \mid \sigma_Z) - s_c\Phi(\theta_0,t,I_p)\big|^\kappa, \qquad \kappa \in \mathbb{N}. \tag{55}$$

Therefore

$$E\big|E(\hat{\Phi}_n(t,I_p) \mid \sigma_Z) - E\hat{\Phi}_n(t,I_p)\big|^\kappa \le 2^{2\kappa} E\big|E(\hat{\Phi}_n(t,I_p) \mid \sigma_Z) - s_c\Phi(\theta_0,t,I_p)\big|^\kappa. \tag{56}$$

To estimate the right side, note that

$$f_{Y,X|Z}(y,x|z) = \iint f_{A,B}(x - a_1 z, a_1, y - c_1 x, c_1)\, da_1\, dc_1\,,$$

so that, by Assumption 12, we have for all $z, z' \in \operatorname{supp} Z$,

$$\int\int_I \big|f_{Y,X|Z}(y,x|z) - f_{Y,X|Z}(y,x|z')\big| dx\, dy \le C_A \cdot |z - z'|\,.$$

Applying this to (53) and (22) yields that

$$\big|E(\hat{\Phi}_n(t,I_p) \mid \sigma_Z) - a_1\Phi(\theta_0,t,I_p)\big| \le 2C_A \sum_{j=1}^{n-1} \big(Z_{(j+1)} - Z_{(j)}\big)^2 + |Z_{(1)} + 1| + |Z_{(n)} - 1|\,, \tag{57}$$

holds true almost surely. Therefore, using Lemma B.1, we get that

$$\begin{aligned}
&E\big|E(\hat{\Phi}_n(t,I_p) \mid \sigma_Z) - E\hat{\Phi}_n(t,I_p)\big|^\kappa \\
&\le 2^{2\kappa} E\Big(2C_A \sum_{j=1}^{n-1} \big(Z_{(j+1)} - Z_{(j)}\big)^2 + |Z_{(1)} + 1| + |Z_{(n)} - 1|\Big)^\kappa \\
&\le 2C_A^\kappa 6^\kappa E\Big[\Big|\sum_{j=1}^{n-1} \big(Z_{(j+1)} - Z_{(j)}\big)^2\Big|^\kappa + E|Z_{(1)} + 1|^\kappa + E|Z_{(n)} - 1|^\kappa \\
&\le 2C_A^\kappa 6^\kappa (n-1)^{\kappa-1} E \sum_{j=1}^{n-1} \big(Z_{(j+1)} - Z_{(j)}\big)^{2\kappa} + E|Z_{(1)} + 1|^\kappa + E|Z_{(n)} - 1|^\kappa \\
&= \mathcal{O}(n^{-\kappa})\,.
\end{aligned} \tag{58}$$

Together with (54) and (52) this implies (50).

To obtain (51), from (54), (57) and Lemma B.1 we get that uniformly in t,

$$\big|E\hat{\Phi}_n(t,I) - s_c\Phi(\theta_0,t,I_p)\big| \le E\big|E(\hat{\Phi}(t,I) \mid \sigma_Z) - s_c\Phi(\theta_0,t,I_p)\big| = \mathcal{O}(1/n)\,.$$

Therefore,

$$\begin{aligned}
&\|E\hat{\Phi}_n(\cdot) - s_c\,\Phi(\theta',\cdot)\|_{v;q} - \|E\hat{\Phi}_n(\cdot) - s_c\,\Phi(\theta_n,\cdot)\|_{v;q} \\
&\ge s_c\,\|\Phi(\theta_0,\cdot) - \Phi(\theta',\cdot)\|_{v;q} - s_c\,\|\Phi(\theta_0,\cdot) - \Phi(\theta_n,\cdot)\|_{v;q} - \mathcal{O}(1/n) \\
&\ge c_{\Theta,0}^{1/2}\|\theta_0 - \theta'\| - c_{\Theta,1}^{1/2}\sqrt{d}bn^{-1/2} - \mathcal{O}(1/n)\,,
\end{aligned}$$

by (48) and Assumption 9. This completes the proof of the theorem. ∎