The Triangular Model with Random Coefficients

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

June 15, 2017

The triangular model is a very popular way to allow for causal inference in the presence of endogeneity. In this model, an outcome is determined by an endogenous regressor, which in turn is first caused by an instrument. We study the triangular model with random coefficients and additional exogenous regressors in both equations, and establish non-identification of the joint distribution of random coefficients. This implies that counterfactual outcomes are not identified either. Non-identification continues to hold if we confine ourselves to the joint distribution of coefficients in the outcome equation or indeed any marginal, except the one on the endogenous regressor. Nonidentification prevails as well, if we focus on means of random coefficients, implying that IV is asymptotically biased. Based on these insights, we derive bounds on the joint distribution of economically relevant functionals, e.g., counterfactual outcomes, and suggest an additional restriction that allows to point identify the distribution of random coefficients in the outcome equation. We extend the model 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, and we provide an application to consumer demand.

Keywords: Random Coefficients, Endogeneity, Nonparametric Estimation, Identification, Characteristic Function, Demand Analysis.

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

[*]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

population differ from a model that neither takes heterogeneity nor endogeneity into account is therefore a question of great importance.

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 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 (2010), Masten (2015), and Masten and Torgovitsky (2015). 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), Graham and Powell (2012), 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), Jun (2009), or Chesher (2003).

What distinguishes our paper from any of the previous contributions, with the notable exception of Masten (2015), 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 take this specification explicitly into account. We focus on high dimensional unobserved heterogeneity, since we believe it to be the most important feature of reality in many applications, while we accept 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 X and the instruments, at the expense of being 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.

A natural question is whether it is necessary to have random coefficients on all economically relevant

variables in the outcome equation, and not just a scalar heterogeneity factor. An important reason for doing so is that the parameters in equation (1) are often interrelated, because they stem from the same decision problem. In the set-up of our application for instance, suppose that an individual maximizes the indirect utility $v(x,w) = (\beta' \ln(w))(\ln(x) - \alpha' \ln(w))$, where x is income, $\ln(w)$ denotes the vector of log prices, and α, β are parameters which are subject to the normalizations $\iota'\beta = 0, \iota'\alpha = 1$. Then, by applying Roy's identity, the resulting demand function for the budget shares of good j takes the form $y_j = \alpha_j + \beta_j(x - \alpha' w) = \alpha_j + \beta_j x + \gamma_j' w$, where $\gamma_j = \beta_j \alpha$. As such, all coefficients are functions of the same deep underlying parameters of the utility function. If the coefficients of the utility functions are heterogeneous across the population, it is thus natural to assume that all coefficients in the outcome equation are random, too, as well as correlated with joint density f_B.

To analyze this model and obtain f_B, 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 most of the literature on random coefficient models (e.g., Hoderlein et al. (2010), Gautier and Kitamura (2013)), we confine ourselves to this setup. Second, as key identifying restriction we will assume full independence of instruments and exogenous covariates from the random coefficients:

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 non-identification, the results would of course continue to hold under weaker forms 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 with smooth densities of compact support. This non-identification result also continues to hold, if we consider the case where Z exerts a monotonic influence on X, 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 we construct several four dimensional densities that are compatible with this aspect of the characteristic function. Moreover, the counterexample shows that one key source of non-identification is related 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 α-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, non-identification extends to the marginal distribution (and indeed the mean) of B_2. Beyond the fact that these distributions are of interest in their own right - think of B_2 in consumer demand for instance as the price effect - it also implies that none of the "joint marginal" distributions are identified, e.g., $f_{B_1 B_2}$. It is thus impossible to obtain the covariances between random parameters, say, between price and income effect (i.e., $Cov(B_1, B_2)$). It is also impossible to identify the distribution of important economic quantities that are functionals of these joint distributions,

e.g., the distribution of welfare effects in consumer demand which, even in the linear model, is a function of both B_1 and B_2 (Hausman (1981)).

These results suggest that we need to impose additional assumptions to identify the joint distribution of random coefficients in the outcome equation f_B, and most marginals. 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 and a school choice application. For applications where this assumption is not considered plausible, we derive bounds on economically important functionals, e.g., the distribution of counterfactual outcomes. In contrast, under this assumption which actually includes the case where there is as little as 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. (2010), see the supplementary material Hoderlein et al. (2016b).

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 semiparametric 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 entirely 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 analyticity of the density of f_B. Since it is of particular relevance for applications, we focus on the former, and show how to construct a 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 estimator in large samples, and show that it achieves a parametric rate of convergence and that it is asymptotically normally distributed. Further, we analyze the behavior of the linear instrumental variables estimator for the means of the random coefficients in the outcome equation, and show its inconsistency. 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. (2010), and Gautier and Hoderlein (2012). It also falls into the wider class of models analyzed in Fox and Gandhi (2015) 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 Masten (2015), who analyzes a fully simultaneous linear random coefficient system, which nests our model. There are several differences, though. Masten (2015) focuses on identification of the marginal distribution of B_1, but he does not provide conditions under which the rest of the model is identified. Moreover, his model does not cover our extended model including W.

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 model, at the expense of identifying (only) the average structural

function, see Imbens and Newey (2009), or some local average structural derivatives, see Hoderlein and Mammen (2007), 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 (2015), who focus on the average random coefficient in a linear correlated random coefficient model with continuous outcome, and Hoderlein and Sherman (2015), who consider the same model with binary outcomes. All of these approaches 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 (2015) and Blundell, Kristensen and Matzkin (2014). See also 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 we show that nonidentification stems at least in parts from nonidentification of EB_0. In Section 2.2 we show that the arguments extend to exogenous covariates. Further, we compute explicitly the limit of linear IV and show that it is asymptotically 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 under an additional independence assumption. An extension that is important for applied work is discussed in Section 3.2: we consider the case of limited support of Z and W. In the absence of the additional identifying assumption, it is at best possible to obtain bounds which are provided in Section 3.3. The point 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 examined through a Monte Carlo study in Section 5. Finally, we apply our estimator for the random coefficients to British consumer data, before an outlook concludes. Proofs are deferred to the appendix, while the supplement Hoderlein et al. (2016a) contains additional technical arguments and material. The supplement Hoderlein et al. (2016b) presents nonparametric estimation theory.

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

The nonidentification results that we present in this section do not rely on infinite support or violation of regularity conditions; indeed, densities can be extremely well behaved, e.g. analytic, and are not pathological at all. Moreover, there is a continuum of densities that are not identified, and hence the counterexamples do not constitute isolated points in a function space. Further, nonidentification continues to 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. Our arguments are nonparametric, meaning that it may be possible to achieve positive identification results through a fully parametric model. However, these would rely exclusively on the parametric assumptions imposed, and would break down in case of a misspecified model. Finally, even if a parametric random coefficients density were identified, our results imply that for every identified density there exists a nonidentified density in any arbitrarily small neighborhood.

In order to study identification we proceed in several steps, and first consider the simplest version of our 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.

2.1. Nonidentification of distribution of the intercept B_0

To understand the nonidentification of B_0, consider the basic model (2). Our counterexample employs the reduced form model, which is obtained by inserting the second equation into the first one, leading to

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

where $C=(C_0,C_1)$, $C_0=B_0+B_1A_0$ and $C_1=B_1A_1$. In order to study the link between the distributions of (A,B) and (A,C) we introduce the mapping $\tau(a_0,a_1,b_0,b_1) := \big(a_0,a_1,b_0+b_1a_0,b_1a_1\big)$. Note that the restriction of τ to the domain $\{a_1\neq 0\}$ represents an invertible mapping to this set with $\tau^{-1}(a_0,a_1,c_0,c_1) = \big(a_0,a_1,c_0-c_1a_0/a_1,c_1/a_1\big)$. Since (A,B) is assumed to have a Lebesgue density $f_{A,B}$ we have that $\{A_1=0\}$ is a null set. It follows that (A,C) has a Lebesgue density $f_{A,C}$ as well and that

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

Recall from Masten (2015) that from (3) the joint density of (A_1,C_1) and, since $B_1=C_1/A_1$, also the joint density of (A_1,B_1) is identified in case of fully supported Z. For B_0 however, this argument fails since the distribution of B_0 cannot be recovered from that of (A_0,C_0), as $C_0=B_0+B_1A_0$, and neither can the distribution of counterfactual outcomes $Y_x=B_0+B_1x$ be identified in this fashion.

In the following, we will formally show nonidentification of these distributions 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 (Y,X) given $Z=z$ to $\psi_{A,C}$ via

$$\begin{aligned} \psi_{X,Y|Z}(t_1,t_2|z) &:= E\big(\exp(it_1X+it_2Y)|Z=z\big) \\ &= E\exp\big(it_1(A_0+A_1z)+it_2(C_0+C_1z)\big) = \psi_{A,C}(t_1,t_1z,t_2,t_2z), \end{aligned} \tag{5}$$

where $z\in\operatorname{supp} Z$. Lemma 10 in the appendix shows that this is actually all the information on (A',C') contained in the distribution of (Y,X,Z).

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

$$\int b_0\eta(a_0,a_1,b_0,b_1)da_0\,da_1\,db_0\,db_1 \neq 0, \tag{6}$$

where $\eta(a_0,a_1,b_0,b_1)$ denotes an infinitely differentiable and compactly supported function, and note that, if η is the difference between two densities, this condition implies that the means $E[B_0]$ of the two densities differ.

Theorem 1. *There exists a density $f_{A,B}(a_0,a_1,b_0,b_1)$ and a function $\eta(a_0,a_1,b_0,b_1)$, both having compact support, such that*

$$f_{A,B;\gamma}(a_0,a_1,b_0,b_1) = f_{A,B}(a_0,a_1,b_0,b_1)+\gamma\eta(a_0,a_1,b_0,b_1), \qquad |\gamma|\leq\gamma_0, \quad \textit{some } \gamma_0>0, \tag{7}$$

is a family of densities and η satisfies (6), and so that the distribution of (Y,X,Z) in model (3) is the same for any density in (7) even if Z has full support. In particular, the mean of B_0 cannot be identified from the distribution of the observations (Y,X,Z).

Remarks. 1. *(Interval of potential values of EB_0).* If we let

$$\bar{b}_0 := \int b_0 f_{A,B}(a_0,a_1,b_0,b_1)da_0\,da_1\,db_0\,db_1,$$

the theorem implies that there is a full interval of potential values of EB_0 with center $\bar{b}_0$, all corresponding to different candidate densities of (A',B') in the family $f_{AB;\gamma}$, all of which generate the same joint distribution of the observables (Y,X,Z). Moreover, as the Appendix A.2 shows, the construction is fully explicit.

2. *(Nonidentification of counterfactual outcomes).* The theorem states that one non-identified element is the marginal of B_0. This is an important quantity in a random coefficients world, because it captures the heterogeneous baseline level of Y, implying that we cannot identify the distribution of $Y_x = B_0 + B_1 x$ for any x in the absence of further assumptions. Hence we cannot identify quantities that are analogous to the quantile treatment effect, i.e., $q_{Y_x}(\alpha) - q_{Y_{x-1}}(\alpha)$, and not knowing f_{B_0} precludes any welfare analysis that is based on the level of Y. For example, if Y was income we could determine the distribution of treatment effects on income, but not whether the treatment has an effect on people that have a low level of income prior to treatment.

3. *(No local identification).* The non-identification result seems to rely on the specific density $f_{A,B}$ used in the counterexample. However, it is straightforward to see that it extends to a much larger class of densities. In particular, start out with an arbitrary four dimensional density f_0, for which the mean of the first coordinate exists. If we consider the mixture $\lambda f_0 + (1-\lambda) f_{A,B;\gamma}$, $|\gamma| \leq \gamma_0$, for any fixed $\lambda \in (0,1)$, by a straightforward extension of our result we obtain that the mean of B_0 is not identified in any neighborhood of any such density f_0. This means that in a neighborhood of any density f_0 (which may or may not be identified) there exists an entire family of non-identified densities which is not identified. Thus, non-identification is pervasive in this setup.

4. *(Intuition behind the result).* The basic intuition behind the theorem is that a three dimensional object, 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. However, in particular to determine EB_0 as a specific source of nonidentification, more involved arguments are required, see Appendix, Section A.1. Finally, as the next section shows, it is not the case that, even in situations where the numbers of dimensions coincides, identification is immediate. Instead, it is a combination of specific structure and enough dimensions of observable variation that ensure identification.

2.2. Nonidentification of distribution of the slope B_2

The results on nonidentification extend to the slope coefficients B_2 in model (1). Apart from making the distributions of counterfactual outcomes $Y_{x,w} = B_0 + B_1 x + B_2' w$ nonidentified, these distributions are of interest in their own right; for example, in our application, B_2 will include the price effect. Further, it is also impossible to identify the distribution of important economic quantities such as welfare effects in consumer demand which, even in the linear model, are a function of both B_1 and B_2 (cf. Hausman (1981)).

For simplicity, in our construction we still assume that there is one excluded instrument Z and add one exogenous covariate W, resulting in a four-dimensional vector (Y,X,Z,W) of observables. If all

coefficients are random and continuously distributed, the vector (A', B') has six dimensions. Thus, it is not so surprising that mapping the four dimensional distribution of observables onto the six dimensional distribution of random coefficients is not possible. Indeed, one source of nonidentification is the mean EB_2, as Theorem 11 in the Appendix, Section A.1 shows. However, it is not just merely counting dimensions. Instead, it is also the nature of the variation that is important. The following special case without intercept, and thus equal dimensionality of RCs and observables, shows that the variation in W does not yield useful additional information.

Theorem 2. *Consider the triangular model (1) under Assumption 1, and suppose that $L = S = 1$ and that $A_0 = B_0 = 0$, and allow (Z,W) to have full support. Then, neither the mean, nor the distribution of B_2 can be identified from the distribution of the observables (Y,X,Z,W). More precisely, if we consider the densities $f_{A,B;\gamma}$ from Theorem 1 as densities for the vector $(A_2,A_1,B_2,B_1)'$, then the distribution of (Y,X,Z,W) in model (1) will be the same for any γ.*

2.3. Linear instrumental variables

Having established that it is impossible to point-identify parameters of interest without further assumptions, a natural question that arises is what standard linear IV identifies and estimates in this model. Stock and Watson (2011), p. 500-501, consider model (2) and show that the second coefficient of linear IV estimates $E[B_1A_1]/EA_1$. We generalize this to the model of the previous section, with one excluded instrument Z and one additional exogenous covariate W. Here, 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 accordingly.

Theorem 3. *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*

$$EA_1\left(EZ^2EW^2 - \left(EZW\right)^2\right) \neq 0, \tag{8}$$

then linear IV estimates the population parameter

$$\mu_{IV} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & E[ZX] & E[ZW] \\ 0 & E[WX] & E\left[W^2\right] \end{pmatrix}^{-1} \begin{pmatrix} EY \\ E[YZ] \\ E[YW] \end{pmatrix} = \begin{pmatrix} EB_0 + E[A_0B_1] \\ E[A_1B_1]/EA_1 \\ EB_2 + E[A_2B_1] - E[A_1B_1]EA_2/EA_1 \end{pmatrix} \tag{9}$$

The proof of this result is deferred to the technical supplement, Hoderlein et al. (2016a), Section E.2.

Remarks. 1. The linear IV estimate is of course asymptotically biased for the nonidentified means of B_0 and B_2, but it is even biased for the identified parameter EB_1. The biases are not signed in general, and depend on correlations of random coefficients across equations. In the next section we introduce the condition that A_1 and $B = (B_0,B_1,B_2)$ be independent as a sufficient condition for point identification of the distribution of B. Under this assumption, μ_{IV} reduces to

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

Then μ_{IV} becomes consistent for EB_1 but remains asymptotically biased for EB_2 if $\mathrm{Cov}(A_2,B_1) \neq 0$. Note, however, that linear IV becomes consistent for EB_2 if A_2 is non-random.

2. The condition (8) requires that the instrument Z has an effect on average ($EA_1 \neq 0$), and that Z and W are not linearly related so that $(EZW)^2 < EZ^2 EW^2$.

3. From (9) we may conclude that the covariance $\mathrm{Cov}(A_2, B_1)$ is not identified either. Indeed, since EA_j, $j = 1,2$, $E[A_1 B_1]$ and μ_{IV} are identified, this follows from nonidentification of EB_2. Therefore, parametric models which rely on first and second moments, in particular the normal distribution, should not be used for the joint distribution of (A', B').

4. The theorem assumes centered regressors and instruments. Not surprisingly, in the general case the bias is substantially more involved. Generally, if we let $\bar{X} = X - EX$, $\bar{W} = W - EW$ and $\bar{Z} = Z - EZ$, then the model can be rewritten as

$$\begin{aligned} Y &= \tilde{B}_0 + B_1 \bar{X} + B_2 \bar{W}, \quad \tilde{B}_0 = B_0 + B_1\, EX + B_2 EW, \\ \bar{X} &= \tilde{A}_0 + A_1 \bar{Z} + A_2 \bar{W}, \quad \tilde{A}_0 = A_0 - EA_0 + (A_1 - EA_1)EZ + (A_2 - EA_2)EW, \end{aligned}$$

Thus, the coefficients A_j and B_j, $j = 1,2$, are left unchanged, and the corresponding conclusions regarding these coefficients remain valid, but the coefficients of the intercepts are more complicated, and the IV estimates based on centered regressors yields $E\tilde{B}_0 + E\tilde{A}_0 B_1$ as IV estimate for the intercept parameter.

3. Identification

After these negative results, it is clear that 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 3.2).

Finally, in Section 3.3 we drop the marginal independence assumption which implies point identification, and present explicit bounds on the distribution of counterfactual outcomes, as well as implicit bounds on general linear functionals.

3.1. Identification of f_B under full support

First we present our result in the basic model (2). Consider the following two assumptions.

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

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.

Discussion of Assumption 3. 1. 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, to provide a structural framework related to our consumer demand application, assume that the first stage equation that determines the choice of total expenditure X, stems from an individual solving an intertemporal consumption model with CARA preferences, i.e., the instantaneous utility function takes the form $u(x;\rho) = -\rho^{-1}\exp(-\rho x)$, and the agent is assumed to face an interest rate $r = R - 1$

(in the following for simplicity riskless and constant). Heterogeneity means, of course, that this preference parameter ρ be allowed to vary across the population. Formally, given information in period t (i.e., $\mathscr{F}_t$), the individual optimizes

$$\max_{(X_{t+\tau})_{\tau=0,..,\tilde{T}-t}} \mathbb{E}_t \left\{ \sum_{\tau=0,...,\tilde{T}-t} \beta^\tau u(X_{t+\tau};\rho) \right\},$$

where $\mathbb{E}_t$ denotes conditional expectation with respect to $\mathscr{F}_t$, β is the discount rate, $\tilde{T}$ is the terminal period (for simplicity assumed to be large), and u is an utility function of the CARA class, taking into account the dynamic constraints

$$\begin{aligned} Q_{t+1} &= (M_t - X_t)R \\ M_{t+1} &= Q_{t+1} + S_{t+1}, \end{aligned}$$

where S_{t+1} is the consumer's idiosyncratic income in period t, M_t denotes nonhuman assets in period t, and total wealth Q_t is generated as above. The most common example in the literature has S_t follow an ARMA process (possibly with a unit root), so for the purpose of this discussion, we choose $S_{t+1} = S_t + \Sigma_{t+1}$, and assume that log income follows a random walk with Gaussian innovations which are *iid* over time, see Deaton (1992). In this case, with σ^2 the variance of the income shock Σ_t, the relationship between log total expenditure and log income becomes

$$X = S + \kappa Q - 0.5\sigma^2\rho,$$

where κ is a proportionality factor (e.g., if $\tilde{T}$ is large, $\kappa \cong \tilde{T}^{-1}$), and we dropped the t subscripts. If we assume that income shocks are heteroskedastic (with higher variance for individuals with a higher income/wage rate Z), we obtain that $\sigma^2 = \gamma Z$. Moreover, we let $S = S_0 + \eta Z$, where η is the yearly hours worked, which at least for the working male population typically does not vary too much, and S_0 is some base level/intercept. Thus, $X = S_0 + \kappa Q + (\eta - 0.5\rho\gamma)Z$. Now, assume as is often the case that assets, including human capital and social security wealth, are not observed in the data, and let this unobservable be denoted by $A_0 = S_0 + \kappa Q$. Moreover, if we allow ρ and γ to be heterogeneous, we arrive exactly at the model $X = A_0 + A_1 Z$, where A_1 reflects heterogeneity in either risk aversion ρ or in the variance η. These factors are likely to be uncorrelated with the coefficients B, which as outlined in the introduction reflect relative preference for a specific good, e.g., food at home. In contrast, the annualized life cycle wealth A_0 is probably the most important part of total expenditure (consumption). Since the broad categories like food analyzed in consumer demand are also important parts of consumption because of the budget constraint (in our data, roughly a quarter to a third), there is likely correlation between A_0 and the unobserved drivers of food consumption (i.e., B), and it is clearly less attractive to assume that there is no correlation between those variables. Another related argument is measurement error (see Lewbel and Nadai (2015)): If food expenditure contain a measurement error, then B_0 will contain a measurement error, too. But since food consumption is a large part of total consumption, the measurement error will also be a part of A_0, and again we have correlation between A_0 and B_0.

Second, consider a workhorse example from labor economics, where the outcome Y is labor income and X is total schooling, including possibly higher education and job related training, assumed to be at least approximately continuously distributed. In this example, B_1 reflects heterogeneity in individuals' wage reactions to additional schooling, which commonly are thought to be largely determined by unobserved work related ability which in turn is correlated to the level of schooling, while B_0 reflects baseline income. In case of the common cost shifter instruments Z, say, distance from college, the heterogeneity in A_1 reflects differences in individuals' schooling levels in response to differences in distance to college. In this example, it is plausible that these heterogeneous effects of distance to college are unrelated to the

work related abilities an individual uses later in life. Conversely, we think of A_0 as reflecting, in the spirit of the Roy model, (expected) benefits of the level of treatment (i.e., schooling), conditional on the individuals' information. Formally, an individual balances $E[Y_x|\mathscr{F}]$, where $\mathscr{F}$ denotes her information and Y_x counterfactual income at level x, with the costs $\tilde{a}_0 + a_1 x$, i.e., she determines the optimal schooling x^* as

$$x^* = \arg\max_{x \in \tilde{X}} \{E[Y_x|\mathscr{F}] - (\tilde{a}_0 + a_1 z)\}.$$

As random variables across the population A_0 hence equals the net benefit $E[Y_x|\mathscr{F}] - \tilde{A}_0$, while A_1 reflects the marginal costs associated with Z, see also Imbens and Newey (2009) for a related model. It seems sensible to assume that the individuals have at least some meaningful unobserved information about both the two components of Y_x, i.e., their base salary B_0 and the effect of schooling, B_1, because otherwise college choice would be random across the population. As such, it seems likely that A_0 be correlated with B_0 and B_1, or, at the very least, significantly more likely than correlation between A_1 and B.

2. 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 expected benefits involving Y. Also, as we will see in the next subsection, once we have several cost factors only the marginal effect of one of these factors need to be independent of B. Further, 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.

Regarding integrability, to have $E|A_1|^{-1} < \infty$ requires $f_{A_1}(a)/|a|$ to be integrable. For this, it suffices to have $f_{A_1}(a) \leq C|a|^\alpha$ for some C, and an $\alpha > 0$ in a neighborhood of 0. In particular, the mass of A_1 in a neighborhood of 0 of length δ must be of smaller order than δ, meaning that the instrument Z is not allowed to affect X too weakly.

In sum, we feel that this assumption is defendable in many applications, however, this should not take away from the fact that it amounts to placing structure on the unobservables - in light of the non-identification results a necessary evil to achieve point identification. ◇

We now give the main steps of the argument which shows identification under Assumptions 1 - 3. In the following we denote by $\mathscr{F}_d$ the d-dimensional Fourier transform. If applied to a one-dimensional conditional density such as $f_{Y|X,Z}(y|x,z)$, we write $\mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z)$.

To start, under Assumption 1 the conditional characteristic function of Y given X and Z equals

$$\begin{aligned}\mathscr{F}_1\left(f_{Y|X,Z}\right)(t|x,z) &= E\left(\exp(itY)|X=x,Z=z\right)\\ &= E\left(\exp\left(it(B_0+B_1x)\right)|A_0+A_1z=x,Z=z\right)\\ &= E\left(\exp\left(it(B_0+B_1x)\right)|A_0+A_1z=x\right),\end{aligned} \tag{10}$$

where the last equality stems from the fact that (A,B) and Z independent implies that (B_0+B_1x, A_0+A_1z) and Z independent for fixed x,z, and hence that Z and B_0+B_1x are independent conditional on A_0+A_1z. It is straightforward to see then that

$$\begin{aligned}&E\left(\exp\left(it(B_0+B_1x)\right)|A_0+A_1z=x\right) f_{X|Z}(x|z)\\ &\quad= \int_{\mathbb{R}^3} \exp\left(it(b_0+b_1x)\right) f_{A_0,A_1,B}(x-a_1z,a_1,b)\,da_1\,db_0\,db_1.\end{aligned} \tag{11}$$

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

$$\begin{aligned}
&\int_{\mathbb{R}} \mathscr{F}_1\big(f_{Y|X,Z}\big)(t|x,z)\, f_{X|Z}(x|z)\, dz \\
&\quad = \int_{\mathbb{R}^4} \exp\big(it(b_0+b_1x)\big)\, f_{A_0,A_1,B}(x-a_1z,a_1,b)\, dz\, da_1\, db_0\, db_1 \\
&\quad = \int_{\mathbb{R}^4} |a_1|^{-1} \exp\big(it(b_0+b_1x)\big)\, f_{A_0,A_1,B}(a_0,a_1,b)\, da_0\, da_1\, db_0\, db_1 \\
&\quad = E\Big(\exp\big(it(B_0+B_1x)\big)\,|A_1|^{-1}\Big).
\end{aligned} \tag{12}$$

This is exactly where Assumption 3 comes into play: it allows to separate out the factor $E\left[|A_1|^{-1}\right]$ from under the expectation, by making it not depend on B. Note also that requiring $E|A_1|^{-1} < \infty$ justifies the existence of the integral at the beginning of (12). Under the additional Assumption 3, we thus obtain that

$$\int_{\mathbb{R}} \mathscr{F}_1\big(f_{Y|X,Z}\big)(t|x,z)\, f_{X|Z}(x|z)\, dz = \big(\mathscr{F}_2 f_B\big)(t,tx)\, E|A_1|^{-1}. \tag{13}$$

To formulate our reconstruction formula, introduce the operator T as

$$\big(Tg\big)(a_0,a_1) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}} \int_{\mathbb{R}} |t| \exp\big(-it\,(a_0+a_1z)\big)\, g(t,z)\, dt\, dz, \tag{14}$$

which is well-defined for functions $g(t,z)$ which satisfy $\int_{\mathbb{R}} \int_{\mathbb{R}} |t|\,|g(t,z)|\, dtdz < \infty$.

Theorem 4. *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

$$f_B(b_0,b_1) = \big(E|A_1|^{-1}\big)^{-1}\, T\Big(\int_{\mathbb{R}} \mathscr{F}_1\big(f_{Y|X,Z}\big)(t|x,z)\, f_{X|Z}(x|z)\, dz\Big)\,(b_0,b_1), \tag{15}$$

where T, see (14), is applied w.r.t. the variables (t,x), and for every $x \in \mathbb{R}$, we have that

$$E|A_1|^{-1} = \int_{\mathbb{R}} f_{X|Z}(x|z)\, dz. \tag{16}$$

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

$$f_B(b_0,b_1) = \big(E|A_1|^{-1}\big)^{-1} \frac{1}{(2\pi)^2} \int_{\mathbb{R}} \int_{\mathbb{R}} |t| \exp(-itb_0)\, \psi_{X,Y|Z}(-tb_1,t|z)\, dt\, dz. \tag{17}$$

Remarks. 1. *(Identification).* 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. 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). This result can be used to construct a nonparametric sample counterpart estimator akin to Hoderlein et al. (2010), see the supplementary material Hoderlein et al. (2016b).

2. *(Testing for endogeneity).* Recall that, if all of A and B are independent, there is no endogeneity. In this case, we also obtain that the distribution of Y given X and Z does not depend on Z. To see this, note

that

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

does not depend on z.

3. *(Non-testability of Assumption 3).* The validity of Assumption 3 can not be tested from the data. Formally, this follows since the counterexample in Theorem 1 can be constructed with the baseline density $f_{A,B}$ having independent coordinates. Then Assumption 3 is satisfied for $\gamma = 0$ in (7) but not for $\gamma \neq 0$. ◇

Now we turn to the extended model (1). The support assumption needs to be modified as follows:

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 need to place conditions on the dependence structure of the random coefficient vector. As it turns out, we only need B to be independent of one of the slope coefficients; indeed, 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 then reads as follows.

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

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 $\int_{\mathbb{R}^{1+K}} |t|^K\, |g(t,v)|\, dt\, dv < \infty$. The following result is a natural extension of Theorem 4 in the special case without W, and a one dimensional Z. In particular, the change of variables step involves only one variable, and this variable appears in the denominator.

Theorem 5. *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 $w\in\mathbb{R}^S$, $z_{-1}\in\mathbb{R}^{L-1}$, $x\in\mathbb{R}$,

$$C^{-1} := E|A_{1,1}|^{-1} = \int_{\mathbb{R}} f_{X|Z,W}(x|z,w)\, dz_1. \tag{18}$$

3.2. Identification in the case of limited support

In applications, it is often not plausible to assume that continuous regressors vary over the whole real line. Therefore, we now extend our approach to deal with compactly supported instruments Z. The first important observation is that we cannot directly follow the approach which led to Theorem 4, since the identifying relation (10) only holds for $(x,z)\in \operatorname{supp}(X,Z)$, which generally does not suffice for integrating out z in the second line of (12). There are two possible routes that follow from this observation.

The first is to limit the support of A ; the second is to invoke assumptions that allow to extrapolate from supp Z. Both strategies have their merits and problems, and both strategies have precedents in the econometric literature. We will now discuss them in turn.

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) with multivariate W, but for simplicity assume to have a univariate Z (so $L = 1$). The support restriction will be as follows:

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

$$\text{supp}\left(\frac{x - A_0 - A_2' w}{A_1}\right) \subseteq \text{supp}\left(Z|W = w\right) =: \mathscr{S}_{Z,w} \tag{19}$$

Discussion of Assumption 6. First, note that since the distribution of A is identified and (X, W') is observed, the relation (19) can in principle be checked by estimating the two supports, which is, however, pretty involved. Below we indicate an alternative approach to identify appropriate values of x and w. To illustrate further that the assumption can be satisfied in reasonable settings, let Z have bounded support in the sense that $\text{supp}\left(Z|W = w\right) = [z_l, z_u]$. Moreover, assume that w is such that $\text{supp}\,(A_0 + A_2' w, A_1) \subset [a_l, a_u] \times [a_{1,l}, a_{1,u}]$, 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

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

To obtain $\text{supp}\left((x - A_0 - A_2' w)/A_1\right) \subset \text{supp}\,(Z|W = w)$ 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 $x \in \text{supp}\,(X|W = w)$ with $a_{1,u} z_l + a_u \le x \le a_{1,l} z_u + a_l$, (19) is satisfied. Hence, since

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

if the support of $Z|W = w$ is sufficiently large as compared to that of $(A_0 + A_2' w, A_1)$, (19) 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 6. *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 $(x, w') \in$ supp (X, W) which satisfy (19), the following holds*

$$\left(\mathscr{F} f_B\right)(t, tx, tw) = \left(E|A_1|^{-1}\right)^{-1} \int_{\mathscr{S}_{Z,w}} \mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x, z, w)\, f_{X|Z,W}(x|z, w)\, dz. \tag{20}$$

Setting $t = 0$ yields in particular that

$$E|A_1|^{-1} = \int_{\mathscr{S}_{Z,w}} f_{X|Z,W}(x|z, w)\, dz \tag{21}$$

Remarks. 1. *(Counterfactual outcomes).* Consider the counterfactual outcome $Y_{x,w} = B_0 + B_1 x + B_2' w$ for $(x, w') \in \text{supp}\,(X, W)$. Since the characteristic function of $Y_{x,w}$ satisfies

$$\Psi_{Y_{x,w}}(t) = \left(\mathscr{F} f_B\right)(t, tx, tw), \tag{22}$$

Theorem 6 shows that the characteristic function and hence the distribution of $Y_{x,w}$ is identified for points (x,w') which satisfy (19). By Fourier inversion, one can obtain an explicit reconstruction formula for the density $f_{Y_{x,w}}$.

2. *(Parametric models for f_B).* 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 typically identifies f_B in a parametric model, such as the normal distribution. Therefore, (20) can (and will be used below) to construct a minimum-distance parametric estimator, which actually achieves the parametric rate under reasonable assumptions.

3. *(Assessing the condition (6)).* Note that equation (20) holds for every x,w, which satisfy the support restrictions. In contrast, the relation (52) in the proof shows that if the support restriction is violated at (x,w'), then the integral in (21) will be smaller than the constant $E|A_1|^{-1}$. Therefore, if we estimate the integral for various values of (x,w'), we may identify regions where it achieves its constant maximal value $E|A_1|^{-1}$, thus identifying appropriate choices of the interval I. See Figure 1 in the simulations. $\diamond$

Analytic Continuation. Now we 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 condition (19) on the random coefficients, by using analytic continuation arguments. While this approach leads to more general identification results, it is difficult to use for actual estimation. In the technical supplement to this paper we outline and analyze a nonparametric estimator based on Theorem 7 below, which has, however, only logarithmic rates of convergence. Moreover, the approach which leads to Theorem 7 cannot be easily used to incorporate a parametric assumption on the density f_B which is desirable from an applied perspective.

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 $C_1 = B_1A_1$ are finite and satisfy

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

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

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

Theorem 7. *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.*

3.3. Partial identification - with an emphasis on policy relevant objects

In this subsection we discuss partial identification in our framework which arises when we drop the additional independence assumptions (i.e., Assumption 5). We focus specifically on policy relevant objects. We start out with explicit bounds that are specific to the distribution of counterfactual outcomes, and then present a general linear programming approach that allows to analyze a multitude of policy relevant objects under a linear programming framework.

Bounds on the distribution of counterfactual outcomes. In this subsection, we provide explicit bounds on the distribution function of the counterfactual outcomes. We consider model (1) with a univariate Z. Write

$$Y_{x,w} = B_0 + B_1 x + B_2' w \in \operatorname{supp}(X, W')$$

for the counterfactual outcome at (x, w'). By following the argument leading to (20), but replacing the exponential by an indicator to be able to obtain meaningful bounds, and adding the factor $g(z)$, we obtain for every $(x, w') \in \operatorname{supp}(X, W')$ for which the support restriction (19) is satisfied that

$$\int_{\mathbb{R}} P\big(Y \le y \,|\, X = x, Z = z, W = w\big)\, g(z)\, f_{X|Z,W}(x|z,w)\, dz = E\Big(1_{Y_{x,w} \le y}\, g\left(\left(x - A_0 - A_2' w\right)/A_1\right)\, |A_1|^{-1}\Big) \tag{23}$$

for any nonnegative function g and $y \in \mathbb{R}$.

Theorem 8. *Consider model (1) with univariate Z. Suppose that $(x, w') \in \operatorname{supp}(X, W')$ satisfies (19). Given a nonnegative function g and $y \in \mathbb{R}$ define $G(y|x,w,g)$ to be the identified object in (23), and let*

$$A_g = g\big((x - A_0 - A_2' w)/A_1\big)\, |A_1|^{-1}.$$

Then we have for the distribution function $F_{Y_{x,w}}(y)$ of $Y_{x,w}$ at y that

$$\begin{aligned} &\sup_{p>1} \sup_{g \ge 0 : E A_g^p < \infty} \big(G(y|x,w,g)^p / E(A_g^p)\big)^{1/(p-1)} \\ \le F_{Y_{x,w}}(y) \le &\inf_{p>1} \inf_{g > 0 : E A_g^{-p} < \infty} \big(G(y|x,w,g)\big)^{p/(p+1)} \left(E(A_g^{-p})\right)^{1/(p+1)}. \end{aligned}$$

Remarks. The bounds in Theorem 8 are based on Hölder's inequality. As a consequence, they are explicit, but cannot be expected to be sharp. Since the function g is allowed to vary, the bounds use the full information contained in the conditional distribution of Y given $X = x$ and $W = w$. The upper and lower bounds themselves are identified if the joint distribution of the coefficients A is identified. This is the case under full support of (Z, W), or in case of compact support if the distribution of A is restricted to the class of distributions which are determined by their moments.

A linear programming approach to obtain bounds for general policy relevant objects. In this subsection we again restrict ourselves to fully supported covariates. Following similar arguments as in our main identification result, we conclude that $E[|A_1|^{-1} \exp(i(t_0 B_0 + t_1 B_1 + t_2' B))\; g((t_1/t_0 - A_0 - A_2' t_2/t_0) A_1^{-1})]$, for any bounded, measurable function g, is identified. Setting $g(z; t_3) = \exp(i t_3 z)$, we thus work with

$$G(t) \equiv E[|A_1|^{-1} \exp(i(t_0 B_0 + t_1 B_1 + t_2' B_2)) \exp(i t_2 (t_1/t_0 - A_0 - A_2' t_2/t_0) A_1^{-1}],$$

which is identified for any $t \in \mathbb{R}^{3+S}$. The basic idea is now to characterize bounds on linear functionals of the joint density $f_{A,B}$ of (A', B') while imposing (linear) restrictions implied by the model. Interesting linear functionals are for instance the distribution of counterfactual outcomes, or certain covariances. Formally, we consider

$$\varphi(f_{A,B}) = E[h(A,B)] = \langle h, f_{A,B} \rangle, \tag{24}$$

where h is a known square-integrable function and $\langle \cdot, \cdot \rangle$ is the inner product.

For each t, let $g_t(a,b) = \exp(i(t_0 b_0 + t_1 b_1 + t_2' b_2 + t_3 (t_1 - t_0 a_0 - t_2' a_2)(t_0 a_1)^{-1}))|a_1|^{-1}$. Then, the identified set of densities $f_{A,B}$ is contained in

$$\mathscr{F} = \{f \in L^2 : f \ge 0,\ \int f = 1,\ \langle g_t, f \rangle = G(t),\ \forall t \in \mathbb{R}^{3+S}\}.$$

The set is characterized by a continuum of linear restrictions on f. Moreover, we know that additional objects are identified, e.g. f_A, see Hoderlein et al. (2010), or $f_{A_1B_1}$, see Masten (2015). We can incorporate this knowledge as additional constraints, i.e., we let $\phi_{0,s_0}(a) = \exp(is_0'a)$ and let $\Phi_0(s_0) = E[\exp(is_0'A)]$, $s_0 \in \mathbb{R}^{2+S}$, and $\phi_{1,s_1}(a_1,b_1) = \exp(i(s_{1,0}a_1 + s_{1,1}b_1))$ and we let the objects that are identified by data be denoted by $\Phi_0(s_0) = E[\exp(is_0'A)]$, $\Phi_1(s_1) = E[\exp(i(s_{1,0}A_1 + s_{1,1}B_1))]$, $s = (s_0', s_1')' \in \mathbb{R}^{4+S}$. Next, we can stack these conditions as $\phi_s = (\phi_{0,s_0}, \phi_{1,s_1})'$ and $\Phi(s) = (\Phi_0(s_0), \Phi_0(s_1))'$. Consider a counterfactual object as in (24). Assume now that the goal is to obtain an upper bound, i.e.,

$$\overline{\varphi} \equiv \sup_f \varphi(f).$$

Then, we set up the linear program to obtain the bound as follows:

$$\begin{aligned}\overline{\varphi} = \sup_{f \in L^2} & \langle h, f \rangle \\ s.t. & f \geq 0, \langle 1, f \rangle = 1 \\ & \langle g_t, f \rangle = G(t),\ t \in \mathbb{R}^{3+2S}, \qquad \langle \phi_s, f \rangle = \phi_{0,s_0},\ s \in \mathbb{R}^{4+S}.\end{aligned}$$

Of course, this is an infinite-dimensional linear program. Whether one can make inference for $\overline{\varphi}$ in practice is a different question, but there is ongoing research about these types of constraints, see e.g. Chernozhukov, Newey and Santos (2015). As is common in this literature, see e.g., Haile and Tamer (2003), the resulting bounds are not uniformly sharp. We leave a more detailed development of these questions to a separate paper.

4. Semiparametric estimation

Next we discuss how the insights obtained from identification 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 supplementary material Hoderlein et al. (2016b), 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 as possible, we develop asymptotic theory only for the simple triangular model (2), but also show how the estimators may be extended to include exogenous covariates W. Throughout, we shall maintain Assumptions 1 and 3.

Our semiparametric estimator will rely on the identification results in Theorem 6. We 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 $c_Z > 0$.

Discussion of Assumption 8. The specific interval $[-1,1]$ is chosen for convenience only. Concerning the lower bound of f_Z, suppose that f_Z is in fact only bounded away from zero on a subinterval K. Then

we may consider the subset of observations (Y_k, X_k, Z_k) for which $Z_k \in K$. This changes the distribution of the $Z's$ to the conditional distribution of Z given that $Z \in K$. Moreover, since the selection is made exclusively based on Z, it is independent of A, B, therefore the resulting observations (Y, X, Z) will have a joint distribution according to model (2) with only the distribution of Z changed, but with the same distribution of the random coefficients (A, B). $\diamond$

We additionally require the support restriction (19) for an interval $I \subseteq \operatorname{supp} X$, i.e. that

$$\operatorname{supp}\left(\frac{x - A_0}{A_1}\right) \subseteq \operatorname{supp} Z = [-1, 1] \quad \forall x \in I. \tag{25}$$

Let us start with the estimation of the scaling factor $E|A_1|^{-1}$. From Theorem 6, 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 $\phi : \mathbb{R} \to (0, \infty)$ with $\operatorname{supp} \phi \subseteq I$ and $\int_I \phi = 1$. Observe that

$$E|A_1|^{-1} = \int_I \phi(x) \int_{-1}^{1} f_{X|Z}(x|z)\, dz\, dx. \tag{26}$$

A sample analogue is the Nadaraya-Watson type estimator

$$\hat{a}_{I,n}^{NW} = \frac{1}{n} \sum_{j=1}^{n} \frac{\phi(X_j)}{\hat{f}_Z(Z_j)} \tag{27}$$

where $\hat{f}_Z$ is a nonparametric estimate for the density f_Z, possibly trimmed away from zero. We propose instead to use the following Priestley-Chao-type estimator

$$\hat{a}_{I,n} = \sum_{j=1}^{n-1} \phi(X_{(j)}) \left(Z_{(j+1)} - Z_{(j)}\right), \tag{28}$$

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

In the appendix we show that $\hat{a}_{I,n}$ is consistent for $E|A_1|^{-1}$ at a parametric rate, and actually asymptotically normally distributed. Compared to the Nadaraya-Watson type approach which has been prominent in the literature so far (see, e.g., Hoderlein et al., 2010), our estimator does not involve additional smoothing parameters or nonparametric estimators. Further, we found in simulations that it outperforms the Nadaraya-Watson-type approach in finite samples, also for the parametric estimators discussed below.

The support of the weight functions ϕ determines the effective expected sample size used in $\hat{a}_{I,n}$, which is at most $nP(X_1 \in I)$. We choose ϕ as the uniform weight on I, other choices with full support on I, which give more mass to the center of I, may be used as well. In our simulations with correctly specified I, there were no substantial differences resulting from the choice of ϕ.

Now 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 possible.

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 (25), and further that Assumption 8 is satisfied. By (20),

$$\int_{-1}^{1} \int \exp(ity) \int_I f_{Y,X|Z}(y,x|z)dx\,dy\,dz = E|A_1|^{-1} \cdot \int_I (\mathscr{F} f_B)(t,tx)dx. \tag{29}$$

Note that this equation holds even in the absence of any functional form assumption. Suppose that $f_B = f_{B,\theta_0}$ belongs to the parametric family of models $\{f_{B,\theta} : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^d$ is a compact set. We estimate the left hand side of (29) 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 $I_1,\ldots,I_q$ be finitely many (distinct) intervals which satisfy (25). For a bounded function $\Phi_1(t,I_j)$, $t \in \mathbb{R}$, $j = 1,\ldots,q$, we set

$$\|\Phi_1(\cdot)\|^2_{\nu;q} := \frac{1}{q} \sum_{j=1}^{q} \int_{\mathbb{R}} \left|\Phi_1(t,I_j)\right|^2 d\nu(t), \tag{30}$$

and note that $\|\cdot\|_{\nu;q}$ defines a seminorm. Let $\hat{a}_{I;n}$ be the estimator for $E|A_1|^{-1}$ in (28). Taking into account (29), we choose our estimator $\hat{\theta}_n$ as a minimizer in $\theta \in \Theta$ of the functional

$$\theta \mapsto \|\hat{\Phi}_n(\cdot) - \hat{a}_{I;n} \cdot \Phi(\theta,\cdot)\|^2_{\nu;q}. \tag{31}$$

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 $I_1,\ldots,I_q$ satisfying (25) and a probability measure ν, such that
1. $\theta \mapsto \|\Phi(\theta,\cdot) - \Phi(\theta_0,\cdot)\|_{\nu;q}$ has a unique zero at θ_0,
2. There are $\varepsilon_0, c_\Theta > 0$ such that for all $\theta \in \Theta$ with $\|\theta - \theta_0\| \le \varepsilon_0$,

$$\|\Phi(\theta,\cdot) - \Phi(\theta_0,\cdot)\|^2_{\nu;q} \ge c^2_\Theta \left\|\theta - \theta_0\right\|^2. \tag{32}$$

The second part of Assumption 9 holds true if the l.h.s. of (32) is, as a function of θ, twice differentiable with a non-singular Hessian matrix at θ_0.

Example (*Bivariate normal distribution: Assumptions and choice of parameters*). Our main example as used in the application will be a normal density f_B. In the technical supplement Hoderlein et al. (2016s), Section E, we give a fully rigorous proof of the validity of Assumption 9 for this most important special case, for which ν has two support points and three disjoint intervals I_j are used. In practice, other choices of the weighting measure are more convenient. In the simulation and the application, we choose ν to be centered Gaussian with standard deviation s. In the application and the simulation, we took $s = 0.1$, but the choice of s did not have strong impact on the resulting minimizers. A closed form expression of the contrast in this case is given in the technical supplement Hoderlein et al. (2016a), Section F. ◇

Under an additional smoothness assumption on the density of (A', B'), Assumption 11 to be found in the appendix, we obtain a parametric rate and asymptotic normality of our estimator.

Theorem 9. *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 compact set. Given Assumptions 1, 3, 8, 9 and 11, the estimator $\hat{\theta}_n$ satisfies*

$$\|\hat{\theta}_n - \theta_0\| = \mathcal{O}_P(n^{-1/2}). \tag{33}$$

Moreover, if in addition f_Z is continuous on $[-1,1]$; if the model $\{f_{B,\theta} : \theta \in \Theta\}$ is such that the function Φ is twice continuously differentiable with respect to θ on the interior of Θ; if all partial derivatives of Φ with respect to θ of order ≤ 2 have finite $\|\cdot\|_{\nu,q}$-norm; if the weight measure ν is chosen with a symmetric density and if θ_0 lies in the interior of Θ, then as $n \to \infty$ we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma) \tag{34}$$

is asymptotically centered normally distributed with the covariance matrix

$$\Sigma = 2(E|A_1|^{-1})^{-2} G^{-1} \Big(\int_{-1}^{1} \sigma^2(z)/f_Z(z) dz \Big) G^{-1},$$

where $\sigma^2(z)$ is defined by (78) in the appendix and where the matrix G is the Gram matrix defined by

$$G := \Big\{ \Big\langle \frac{\partial}{\partial \theta_j} \Phi(\theta_0, \cdot), \frac{\partial}{\partial \theta_{j'}} \Phi(\theta_0, \cdot) \Big\rangle_{\nu,q} \Big\}_{j,j'=1,\ldots,d}.$$

Remark. (Asymptotic variance and bootstrap). The asymptotic covariance matrix in (34) is fairly involved. Although it is in principle possible to replace it by an estimator, we rather recommend to use the bootstrap. The smooth bootstrap seems to be well suited for our problem. Here, one adds a small jitter to the sample obtained from the n out of n bootstrap, which formally means that one samples from a smoothed version of the empirical distribution function, see e.g. Silverman (1981). Establishing consistency of the smooth bootstrap requires to go through all the non-standard arguments which lead to (34), but replacing the true density of (Y_k, X_k, Z_k) by a density estimate from which the bootstrap sample is obtained. The details are beyond the scope of the present paper, but similar techniques for the smooth bootstrap were used in Mammen et al. (1993). In the simulations we illustrate that the smooth bootstrap seems to perform well in practice. ◇

Remark. (Nonparametric estimation). The fully nonparametric estimators for the density of f_B that we present in the supplementary material Hoderlein et al. (2016b) are based on the identification results in Theorem 4 (full support, kernel estimator) and Theorem 7 (limited support, analytic extension). Specifying the marginal distribution of A parametrically cannot be taken advantage of in these approaches. However, for $x \in I$, satisfying the support restriction (25), one can use Theorem 6 to estimate the density of the counterfactual outcome $Y_x = B_0 + B_1 x$ nonparametrically. Indeed, observing (22) we obtain the inversion formula

$$f_{Y_x}(y) = \frac{1}{(2\pi) E|A_1|^{-1}} \int_{\mathbb{R}} \int_{-1}^{1} \int_{\mathbb{R}} e^{-it(y-\tilde{y})} \frac{f_{Y,X,Z}(\tilde{y}, x, z)}{f_Z(z)} \, dt \, dz \, d\tilde{y}.$$

By regularizing the inner Fourier inversion, localizing at x and using the Priestly-Chao technique to estimate the integral over the variable z one can obtain a nonparametric estimator of f_{Y_x}. ◇

Remark. (Extensions to exogenous covariates W). 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. Moreover, we assume that the support $\mathscr{S}_{Z,w} = I_Z$ 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 (19) for a rectangle $I_X \times I_W \subseteq \operatorname{supp}(X,W)$, i.e.

$$\operatorname{supp}\left(\frac{x - A_0 - wA_2}{A_1}\right) \subseteq \operatorname{supp}\big(Z|W=w\big) = I_Z\,, \quad \forall\,(x,w)' \in I_X \times I_W\,. \tag{35}$$

In addition, 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 $c > 0$. Recall that from Theorem 6, for all $(x,w)' \in I_X \times I_W$, and $E|A_1|^{-1} = \int_{-1}^{1} f_{X|Z,W}(x|z,w)\,dz$. Choosing a bounded weight function $\phi : \mathbb{R}^2 \to (0,\infty)$ with $\operatorname{supp}\phi \subseteq I_X \times I_W$ and $\int\int \phi = 1$, we hence deduce that

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

For estimation, we recommend to use the following Priestly-Chao type weights which generalize the weights $Z_{(j+1)} - Z_{(j)}$ from the scenario without W,

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

$j = 1,\ldots,n$, 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

$$\hat{a}_{I_X \times I_W,n} = \sum_{j=1}^{n} \phi\big(X_j,W_j\big)\,\lambda_{j,PC}. \tag{36}$$

Nadaraya-Watson-type weights of the form $\big(n\,\hat{f}_{X,W}(X_j,W_j)\big)^{-1}$ are also possible. These weights require an extra choice of bandwidths involved in the bivariate kernel estimator. Further, we found in simulations that even though computation of $\lambda_{j,PC}$ is quite challenging, the performance of these weights is much superior to the Nadaraya-Watson-type weights, indeed even more so than in the situation without covariate W. The contrast is then constructed similarly as above, see the technical supplement Hoderlein et al. (2016a), Section G, for further details.

5. Simulation Study

We investigate the finite-sample performance of the semiparametric estimators of Section 4 in a simulation study. For brevity we only report results of a simulation in the simple model (2). Additional simulations results in the extended model (1) are presented in the technical supplement Hoderlein et al. (2016a), Section G.2.

Data generating process

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

- $A_0 \sim U(0,3)$, and

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

- 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 $\rho_{B_0,B_1} = 0.7$).
- $Z \sim 12 \cdot \text{Beta}(0.5, 0.5)$.

Determining the interval I in (25)

For this data-generating process, the maximal interval that satisfies (25) is $I = [3,12]$, see the example following (25). The proportion of X-values that falls into I is about 44% (large-scale simulation), thus the effective sample size is 0.44 times the actual sample size when using the full interval I.

As discussed below Theorem 6, to identify an appropriate interval, we may estimate the integral $\int_{\mathscr{S}_Z} f_{X|Z}(x|z)\,dz$ for individual x. Then the interval will consist of those x for which the maximal value, which must be equal to $E|A_1|^{-1}$, is attained. As an estimator for individual x, we take

$$\hat{a}_n(x;h) = \sum_{j=1}^{n-1} K_h\big(x - X_{(j)}\big)\,\big(Z_{(j+1)} - Z_{(j)}\big),$$

where K is a non-negative, symmetric, bounded kernel ($\int K = 1$), $h > 0$ a bandwidth and $K_h(x) = K(x/h)/h$. In Figure 1 we plot the corresponding estimates using the Gaussian kernel and a bandwidth of $h = 1$, with total sample size 10^5. From the plot one could choose the interval $[4.2, 13]$. This is slightly misspecified (since it is not contained in $[3,12]$). Below we conduct simulations with the correct interval $I = [3,12]$ as well as with the slightly misspecified interval $I = [4.2,13]$.

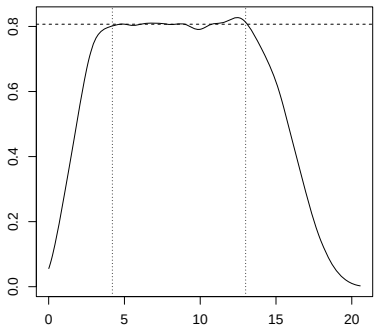


Figure 1: Local estimates of $E|A_1|^{-1}$. Horizontal line: true value 0.8065, vertical lines at $x = 4.2$ and $x = 13$

Estimation of the scaling constant $E|A_1|^{-1}$

First, we consider the estimator (28) of the scaling constant $E|A_1|^{-1}$ from the first-stage equation $X = A_0 + A_1 Z$. Numerical evaluation of $E|A_1|^{-1}$ gives ≈ 0.8065. We take ϕ equal to the uniform density over the intervals $I = [3,12]$ or $[4.2,13]$, respectively. Simulation results for $m = 10000$ iterations and sample sizes $n \in \{500, 1000, 10000, 20000\}$ are summarized in Table 1. In case of the misspecified interval $I = [4.2,13]$, there seems to be a very small bias of order $2 \cdot 10^{-4}$, and the variance is slightly

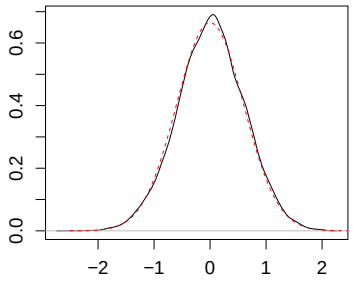


Figure 2: Density estimate of $\sqrt{n}\big(\hat{a}_{I,n} - E|A_1|^{-1}\big)$ in $m = 10000$ iterations for $n = 1000$ (black solid line) and density of $N(0, 0.36)$ (dashed red line)

n	Interval I	Bias	Std	MSE	Total SE
500	$I = [3, 12]$	$13 \cdot 10^{-4}$	0.027	$7.3 \cdot 10^{-4}$	0.36
1000		$8 \cdot 10^{-4}$	0.018	$3.5 \cdot 10^{-4}$	0.35
10000		$0.5 \cdot 10^{-4}$	0.006	$0.36 \cdot 10^{-4}$	0.36
20000		$0.2 \cdot 10^{-4}$	0.004	$0.18 \cdot 10^{-4}$	0.36
500	$I = [4.2, 13]$	$3 \cdot 10^{-4}$	0.028	$8.1 \cdot 10^{-4}$	0.41
1000		$3 \cdot 10^{-4}$	0.020	$4.0 \cdot 10^{-4}$	0.40
10000		$3 \cdot 10^{-4}$	0.006	$0.4 \cdot 10^{-4}$	0.40
20000		$2 \cdot 10^{-4}$	0.004	$0.2 \cdot 10^{-4}$	0.40

Table 1: Statistics for estimator of $E|A_1|^{-1}$ and intervals $I = [3, 12]$ and $I = [4.2, 13]$

increased due to the smaller number of observations in this interval. Generally, the results are very much comparable to those for $I = [3, 12]$.

Concerning asymptotic normality, in case $I = [3, 12]$ the asymptotic variance constant in Proposition 17 is numerically evaluated as 0.36, just as Total SE in Table 1. An oracle version of the Nadaraya-Watson estimator (27), where we replace the denominator by the true density of Z, has asymptotic variance $\mathrm{var}\big(\phi(X)/f_Z(Z)\big)$, which in our setting is numerically evaluated as 0.87. Thus, our estimator is more efficient than the oracle Nadaraya-Watson version. Figure 2 depicts the nonparametric density estimate of $\sqrt{n}\big(\hat{a}_{I,n} - E|A_1|^{-1}\big)$ using the $m = 10000$ iterations for sample size $n = 1000$, together with the normal $N(0, 0.36)$ density.

Finally, Table 2 contains results of bootstrap estimates of the asymptotic standard deviation 0.6 for samples of sizes $n = 1000$ and $n = 10000$. We performed $m = 1000$ bootstrap iterations within each of 1000 iterations. The parameter $h_0 = 0$ indicates the simple n-out-of-n bootstrap, the other values h_i of the bandwidth correspond to the standard deviation in the normal kernel when using the smooth bootstrap. The estimates from the ordinary bootstrap are reasonable, but some improvement can be obtained from the smooth bootstrap, although the choice of h is apparently an issue.

Estimating the parametric model

Next, we consider estimation of the parameters of the normal distribution of (B_0, B_1). In the contrast function (30), we choose the weighting measure $d\nu(t)$ to be centered Gaussian with standard deviations

n		$h_0 = 0$	$h_1 = 0.03$	$h_2 = 0.1$	$h_3 = 0.2$
1000	mean	0.55	0.60	0.60	0.62
	sd	0.03	0.02	0.02	0.02
10000	mean	0.55	0.60	0.60	0.65
	sd	0.02	0.02	0.02	0.02

Table 2: Statistics of bootstrap estimates of the asymptotic standard deviation of $\hat{a}_{I,n}$ using $I = [3,12]$, true value at 0.6, for distinct choices of bandwidth h_i in smooth bootstrap and sample sizes. We used $m = 1000$ bootstrap iterations, and 1000 repetitions in the simulation.

$s = 0.1$ and $s = 0.6$. An explicit form of the contrast is presented in the supplementary material. We partition the interval $I = [3,12]$ and $I = [4.2,13]$ 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, using $I = [3,12]$ and $s = 0.1$. 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 in line with our theory.

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 function 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 function. 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.

Repeatedly performing this algorithm for various sample sizes, using $I = [3,12]$ and $s = 0.1$ we obtain the results in Tables 3 and 4. The estimates of all parameters are quite reasonable. The MSE for estimating μ_0 is much higher than for μ_1. This is 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 appear to have a small bias. Further, in Tables 5 and 6 we present some results for estimates of μ_0 and σ_0 when using $I = [4.2,13]$ or using $s = 0.6$. The results in Table 5 for μ_0 are all rather similar, but the estimator for σ_0 (and similarly for σ_1 and ρ) seems to be somewhat more sensitive to the choice of s. Finally, Table 7 contains results of bootstrap estimates of the standard deviation for the parameters μ_0 and σ_1 for distinct bandwidths. Although less precise than for the scaling constant, the estimates are still reasonable.

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

Table 3: Statistics for estimates of coefficients μ_0 and μ_1, using $I = [3,12]$ and $s = 0.1$

N		Mean	Bias	Std	MSE	Total SE
2000	σ_0	2.048	0.048	0.034	0.003	6
	σ_1	1.009	0.009	0.131	0.017	34
	ρ	0.647	-0.053	0.156	0.027	54
5000	σ_0	2.047	0.047	0.025	0.003	15
	σ_1	0.998	-0.002	0.093	0.009	45
	ρ	0.650	-0.050	0.142	0.023	115
10000	σ_0	2.049	0.049	0.013	0.003	30
	σ_1	1.002	0.002	0.071	0.005	50
	ρ	0.664	-0.036	0.009	0.001	10
20000	σ_0	2.050	0.050	0.012	0.003	60
	σ_1	0.996	-0.004	0.063	0.004	80
	ρ	0.662	-0.038	0.048	0.004	80

Table 4: Statistics for the coefficients σ_0,σ_1 and ρ, using $I = [3,12]$ and $s = 0.1$

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[1]. 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

[1]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).

n	Interval	Param. s	Mean	Bias	Std	MSE	Total SE
2000	[3,12]	0.1	5.06	0.06	0.48	0.24	471
	[4.2,13]	0.1	5.09	0.09	0.53	0.28	567
	[3,12]	0.6	5.08	0.08	0.47	0.22	450
10000	[3,12]	0.1	5.00	-0.01	0.23	0.05	526
	[4.2,13]	0.1	5.01	0.01	0.26	0.07	694
	[3,12]	0.6	5.01	0.01	0.23	0.06	552

Table 5: Statistics for the coefficients μ_0 for distinct choices of parameters used in estimators

n	Interval	Param. s	Mean	Bias	Std	MSE	Total SE
2000	[3,12]	0.1	2.05	0.05	0.03	0.00	6
	[4.2,13]	0.1	2.03	0.03	0.03	0.00	4
	[3,12]	0.6	1.97	-0.03	0.16	0.03	53
10000	[3,12]	0.1	2.05	0.05	0.01	0.00	30
	[4.2,13]	0.1	2.04	0.04	0.01	0.00	14
	[3,12]	0.6	1.94	-0.06	0.17	0.030	337

Table 6: Statistics for the coefficients σ_0 for distinct choices of parameters used in estimators

function of observable demographic variables W_i and a time variable T_i as well, and for all coefficients b_i to in addition vary across the populations, leading the overall model

$$BSF_i = B_{0i} + B_{1i}\ln(TotExp_i) + B_{2i}\ln(Foodprice_i) + b_3 W_{1i} + b_4 W_{2i} + b_5 T_i,$$

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

$$\ln(TotExp_i) = A_{0i} + A_{1i}\ln(Income_i) + A_{2i}\ln(Foodprice_i) + a_3 W_{1i} + a_4 W_{2i} + a_5 T_i,$$

and the standard argument for the validity of income as an IV is that for the type of households we consider (two person households, no children), labor supply is rather inelastic and variations in labor

n	param.		$h_0 = 0$	$h_1 = 0.1$	$h_2 = 0.2$	$h_3 = 0.3$	sim. value
2000	μ_0	mean	19.6	20.9	21	21.4	21.7
	μ_0	sd	2.0	1.7	1.7	1.7	
	σ_1	mean	5.5	5.8	5.7	6.0	5.8
	σ_1	sd	2.0	1.8	1.8	1.9	
10000	μ_0	mean	21.3	23.0	23.4	24.1	22.9
	μ_0	sd	1.4	1.0	0.9	0.9	
	σ_1	mean	8.2	8.3	8.5	8.8	7.1
	σ_1	sd	1.4	1.3	1.2	1.2	

Table 7: Statistics of bootstrap estimates of the asymptotic standard deviation of of μ_0 and σ_1 using $I = [3,12]$, for distinct choices of bandwidth h_i in smooth bootstrap and sample sizes. We used $m = 1000$ bootstrap iterations, and 1000 repetitions in the simulation.

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 in food expenditure, which in turn account for some of the endogeneity in total expenditure. We control again for observable household characteristics through the principal components and include a time trend.

Beyond independence of all random coefficients from the exogenous variables (prices, household characteristics), our main additional assumption specifies A_{1i} to be independent of B_i, which we motivated in Section 3.1. In addition, to avoid being entirely nonparametric, we specify B to have a Normal distribution.

6.2. The Data: The British Family Expenditure Survey

The FES reports several yearly cross sections 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} and 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. Table 8 provides some descriptive statistics of the data, for additional descriptive statistics we refer to Hoderlein (2011), who uses very similar data.

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 $\widetilde{BSF}_i = BSF_i - \left[\widehat{b}_3 W_{1i} + \widehat{b}_4 W_{2i} + \widehat{b}_5 T_i\right]$,

	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

Table 8: Some descriptive statistics of the consumer demand data from the FES

and $\ln(\widetilde{TotExp_i}) = \ln(TotExp_i) - [\widehat{a}_3 W_{1i} + \widehat{a}_4 W_{2i} + \widehat{a}_5 T_i]$, 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 coefficients on the same variables are not random. We therefore arrive at the following 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 $W_{0i} = \ln(Foodprice_i)$.

We determined appropriate intervals of the supports of the X_i's and W_i's by looking for stable regions of local estimates of the scaling constant, similar to Figure 1 in Hoderlein et al. (2016a). We chose $[4.4, 6.0]$ as interval for the X_i, and the full range for the W_i. 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. As starting values we use the IV estimates.

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 $\hat{E}B_j$ denotes either the estimate of the mean computed in the previous iteration or the IV estimates in 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

Let us summarize and interpret the resulting estimates. The estimates for the mean parameters, together with accuracy measures obtained from the simple n-out-of-n bootstrap are given in Table 9.

	IV est.	Semiparam. est.	Std.	.275 Quantile	.975 Quantile
E(B_0)	0.6330	0.6331	0.0199	0.5975	0.6706
E(B_1)	-0.0999	-0.0999	0.0037	-0.1075	-0.0934
E(B_2)	-0.0842	-0.2598	0.1102	-0.4235	-0.0103

Table 9: Estimates of mean coefficients. Preliminary IV estimates and final semiparametric estimates together with estimates of standard errors and quantiles from n-out-of-n bootstrap

The means are generally precisely estimated, and of a very reasonable 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 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 95% of the prices are between -0.07 and 0.07, this means that the budget share of a person with average price semi-elasticity is 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 have no movement in the former coefficient, while the second nearly triples. In fact, price effects are only of a sizeable magnitude 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, see Table 10. In particular, there seems to be 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.

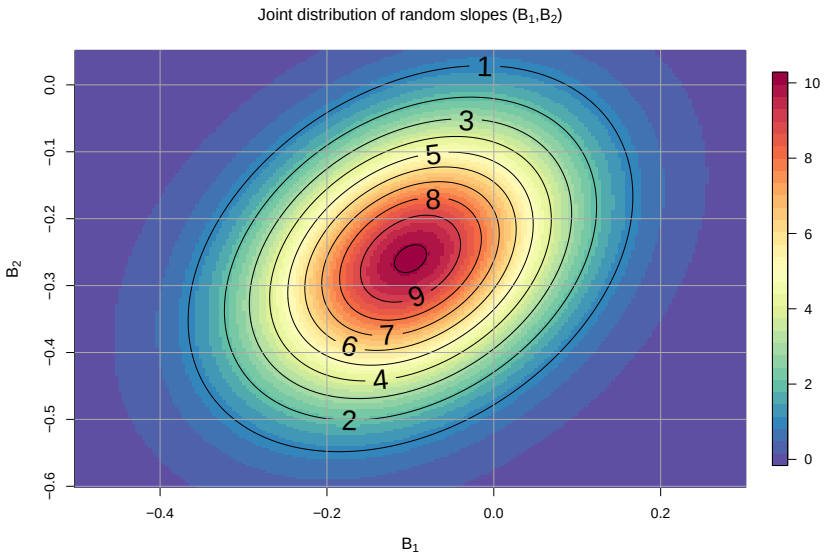
	estimate	Std.		estimate	Std.
Var(B_0)	0.4712	0.0641	Cov(B_0, B_1)	-0.0850	0.1050
Var(B_1)	0.0153	0.0225	Cov(B_0, B_2)	-0.0297	0.0276
Var(B_2)	0.0180	0.0066	Cov(B_1, B_2)	0.0053	0.0064

Table 10: Estimates of the variances and covariances. Semiparametric estimates together with estimates of standard errors from n-out-of-n bootstrap

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, and show a 2D plot of the marginal density of the two slope coefficients B_1 and B_2 in Figure 3, which is to be interpreted as a geographical map involving lines of similar altitude.

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 open question is whether this effect is due to the parametric nature of our approach and whether there truly are parts of the population with positive income effects. We leave this question, which involves a more nonparametric approach, for future research. However, given the pronounced uncertainty associated with the variance parameters, we feel that it is questionable whether there are any individuals whose marginal total expenditure effects are positive and sizeable (say, 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 expendi-

Figure 3: Marginal distribution of the slope on log total expenditures (B_1) and log food prices (B_2) length of X interval is 1.6.



ture and preferences may well be correlated). 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. All of these results have strong implications for welfare analysis, as the welfare effects are largely built on both coefficients.

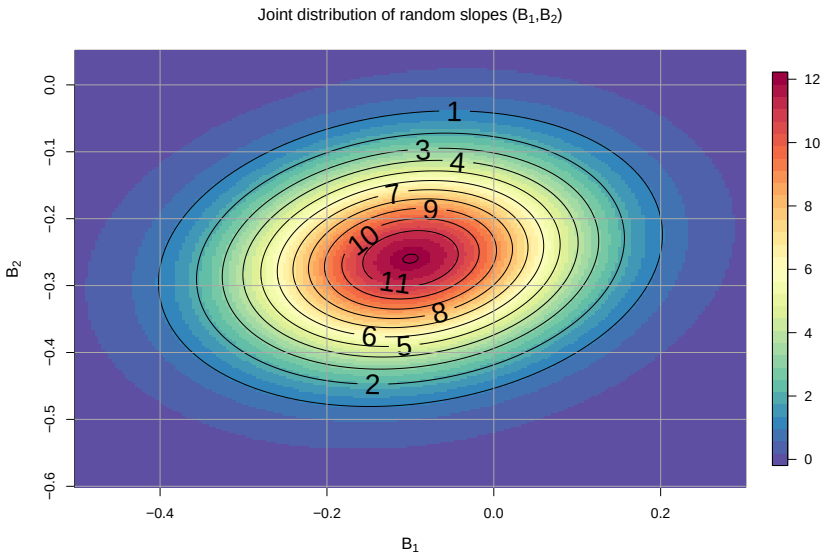
Finally, the last figure shows the same joint density, but for a different value of the interval width. The results are somewhat, but not overly, sensitive to the choice of this parameter; qualitatively they remain preserved. Not surprisingly, we find the stronger changes in B_2 dimension, reflecting perhaps the fact that the price variation is not very plentiful, and hence the estimates are less reliable.

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

Figure 4: Marginal distribution of the slope on log total expenditures (B_1) and log food prices (B_2) length of X interval is 1.7.



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,* **79**, 527-539.

[2] BERAN, R. and P. HALL (1992). Estimating Coefficient Distributions in Random Coefficient Regressions, *Annals of Statistics, **20*** , 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*, **24**, 2569–2592.

[4] BLUNDELL, R., KRISTENSEN, D., and R. MATZKIN (2014). Bounding Quantile Demand Functions using Revealed Preference Inequalities, *Journal of Econometrics*, **179**, 112-127.

[5] BLUNDELL, R., PASHARDES, P. and G. WEBER (1993). What do we Learn About Consumer Demand Patterns from Micro Data?, *American Economic Review*, **83**, 570-597.

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

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

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

[9] CHERNOZHUKOV, V., NEWEY, W., and A. SANTOS (2015). Constrained Conditional Moment Restriction Models, Working Paper, MIT.

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

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

[12] 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*, **76**, 1191-1206

[13] FOX, J. and A. GANDHI, (2009). Full Identification in the Generalized Selection Model, Working Paper, Rice University.

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

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

[16] GRAHAM, B. and J. POWELL (2012). Identification and Estimation of Average Partial Effects in "Irregular" Correlated Random Coefficient Panel Data Models, *Econometrica*,**80**, 2105-2152.

[17] HAILE, P. AND E. TAMER (2003). Inference with an Incomplete Model of English Auctions, *Journal of Political Economy*, 1-51.

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

[19] HAUSMAN, J. and W. NEWEY, (2015). Individual Heterogeneity and Average Welfare, *Econometrica*, forthcoming.

[20] 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.

[21] HODERLEIN, S. (2011). How Many Consumers are Rational, *Journal of Econometrics*, **164**, 294–309.

[22] HODERLEIN, S, H. HOLZMANN and A. MEISTER (2016a). Technical Supplement for: The Triangular Model with Random Coefficients, *Working Paper, Marburg University*.

[23] HODERLEIN, S, H. HOLZMANN and A. MEISTER (2016b). The Triangular Model with Random Coefficients: Nonparametric estimation theory, *Working Paper, Marburg University*.

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

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

[26] HODERLEIN, S. and B. SHERMAN (2015). Identification and Estimation in a Correlated Random Coefficients Binary Response Model, *Journal of Econometrics*, **188**, 135-149.

[27] IMBENS, G. and W. NEWEY (2009). Identification and Estimation of Triangular Simultaneous Equations Models without Additivity, *Econometrica*, **77**, 1481-1512.

[28] JUN, S., (2009). Local Structural Quantile Effects in a Model with a Nonseparable Control Variable, *Journal of Econometrics*, 151, 82-97.

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

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

[31] LEWBEL, A. and K. PENDAKUR (2014). Unobserved Preference Heterogeneity in Demand Using Generalized Random Coefficients, Working Paper, Boston College.

[32] MAMMEN, E., MARRON, J. S. and FISHER, N. I. (1993). Some asymptotics for multimodality tests based on kernel density estimates, *Probability Theory and Related Fields*, **91**, 115-132.

[33] MASTEN, M. (2015). Random Coefficients on Endogenous Variables in Simultaneous Equation Models, Working Paper, Duke University.

[34] MASTEN, M. and A. TORGOVITSKY (2015). Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model, *Review of Economics and Statistics,* forthcoming.

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

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

[37] SILVERMAN, B. (1981). Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society, Series B*, **43**, 97-99.

[38] STOCK, J. H. and M. W. Watson (2011). *Introduction to Econometrics*.

[39] VAN DER VAART, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

[40] 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

Lemma 10. *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 $(\tilde{Y},\tilde{X},Z)$ 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*

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

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

Proof of Lemma 10. 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 $z \in \operatorname{supp} Z$, or, equivalently, that the conditional characteristic functions coincide, which immediately follows from (5) and the assumption in the lemma. ∎

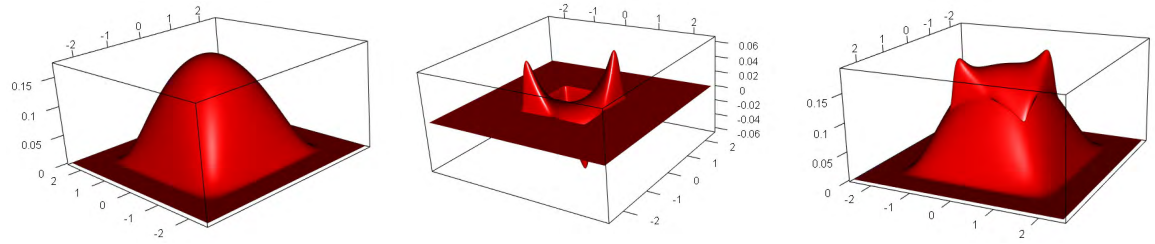


Figure 5: Bivariate marginal density of (A_0, C_1) in the counterexample to identification. Left: Marginal under joint density G, Middle: the function r from the counterexample, Right: Marginal under joint density $G+r$.

Recall that we let $\mathscr{F}_d$ denote the d-dimensional Fourier transform.

Before turning to the details of the proof of Theorem 1, we give an outline concerning the main steps. Start out by noticing that for the polynomial $Q(u_0,u_1,v_0,v_1) = u_0v_1 - u_1v_0$, we have that

$$\mathscr{S} \subset \big\{(u_0,u_1,v_0,v_1) \in \mathbb{R}^4 : \quad Q\big(-i(u_0,u_1,v_0,v_1)\big) = 0\big\}. \tag{37}$$

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 homogenous of degree two, we have $Q\big(-i(u_0,u_1,v_0,v_1)\big) = -Q(u_0,u_1,v_0,v_1)$.

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

$$\begin{aligned}\big(\mathscr{F}_4 r\big)(u_0,u_1,v_0,v_1) &= \Big(\mathscr{F}_4\big[Q\big(\partial_{a_0},\partial_{a_1},\partial_{c_0},\partial_{c_1}\big)G_1\big]\Big) \\ &= Q\big(-i(u_0,u_1,v_0,v_1)\big)\,\big(\mathscr{F}_4 G_1\big)(u_0,u_1,v_0,v_1).\end{aligned} \tag{38}$$

By (37), this implies that the Fourier transform of the function r vanishes on the set $\mathscr{S}$, and since $0 \in \mathscr{S}$, it follows 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 transforms coincide with that of G on the set $\mathscr{S}$. By Lemma 10, 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), where the function η results from r, so that this joint distribution is not identified as well. Figure 5 illustrates this construction. The left hand side corresponds to the two-dimensional marginals of G which is a product density. In the center, the function r is plotted. Finally, on the right we see the marginal of (A_0,C_1) (and (A_1,C_0)) under the joint density $G+r$ for (A,C), the other two-dimensional marginals are equal to those under G itself.

To show that neither marginal of B_0, nor the mean EB_0 is identified is considerably harder. The main step in the nonidentification of EB_0 boils down to showing (see (4)) that

$$\int_{\mathbb{R}} b_0 \int_{\mathbb{R}^3} |a_1|\, r\big(a_0,a_1,b_0+a_1b_0,a_1b_1\big)\, da\, db_1\, db_0 \neq 0, \tag{39}$$

since the difference of the densities is a multiple of r. This is accomplished by using arguments from Fourier analysis, see below.

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. First, we introduce the functions $f_{A,B}(a,b)$, $\eta(a,b)$ and the parameter $\gamma_0 > 0$.

Consider the density on the line

$$g_\beta(s) = \alpha\beta \cdot \exp\big(1/(\beta^2 s^2 - 1)\big)\, 1_{(-1,1)}(\beta s), \quad s \in \mathbb{R}, \tag{40}$$

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. Form 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.

We let

$$f_{A,C}(a,c) = G_{1/2}\big((a,c) - (0,3,0,0)\big)$$

which has support in $[-2,2] \times [1,3] \times [-2,2]^2$, and using (4), let

$$f_{A,B}(a,b) = f_{A,C}\big(\tau(a,b)\big)\, a_1,$$

with support in $[-2,2] \times [1,3] \times [-6,6] \times [-4,4]$. Consider the non-constant polynomial $Q(u,v) = u_0 v_1 - u_1 v_0$, $(u,v) \in \mathbb{R}^4$ as introduced above, and recall that for the set $\mathscr{S}$ in Lemma 10 we have the relation (37). Set

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

Then we have (38) as observed above, so that $\mathscr{F}_4 r$ vanishes on $\mathscr{S}$. Set

$$\tilde{r}(a,c) = r\big((a,c) - (0,3,0,0)\big),$$

so that

$$\big(\mathscr{F}_4 \tilde{r}\big)(u,v) = \exp(3iu_1)\big(\mathscr{F}_4 r\big)(u,v) \tag{41}$$

still vanishes on $\mathscr{S}$, and finally we set

$$\eta(a,b) = \tilde{r}\big(\tau(a,b)\big)\, a_1.$$

Since $Q(0) = 0$, $0 \in \mathscr{S}$, the integrals (corresponding to the Fourier transform at 0) of r, $\tilde{r}$ and η vanish. Further, since r is continuous with support in $[-1,1]$, and $G_{1/2}$ is continuous and bounded away from 0 on $[-1,1]$, there is a $\gamma_0 > 0$ such that $G_{1/2} + \gamma r$ is still non-negative for all $|\gamma| \le \gamma_0$, so that

$$f_{A,C;\gamma} = f_{A,C} + \gamma\tilde{r}$$

and hence also $f_{A,B;\gamma} = f_{A,B} + \gamma\eta$ are densities. From Lemma 10, we get that the distribution of the observed (Y,X,Z) coincide for all γ.

It remains to show that (39), that is,

$$\int_{\mathbb{R}^4} b_0\, \eta(a_0,a_1,b_0,b_1)\, da\, db \neq 0.$$

Step 2. As an intermediate step, we show the general formula

$$f_B(b) = \frac{-i}{(2\pi)^2} \int_{\mathbb{R}^2} \exp(-iv_0 b_0) \frac{\partial \psi_{A,C}}{\partial u_1}(-v_0 b_1, -v_1 b_1, v_0, v_1)\, dv. \tag{42}$$

for the joint density f_B of B, if

$$\int_{\mathbb{R}^4} 1_{(-\infty,0]}(a_1) f_{A,C}(a,c)\, da\, dc = 0, \tag{43}$$

and if

$$\int_{\mathbb{R}^2} \sup_{u \in \mathbb{R}^2} \big|\partial_{u_1} \psi_{A,C}(u,v)\big| dv < \infty. \tag{44}$$

To show (42), 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 (43) we get that

$$f_B(b) = \int f_{A,B}(a,b)da = \int f_{A,C}\big(\tau(a,b)\big)|a_1|\,da = \lim_{\delta\downarrow 0}\int K(a\delta) f_{A,C}\big(\tau(a,b)\big)a_1\,da\,. \tag{45}$$

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 (45) and letting $\delta\to 0$ and using dominated convergence, which is justified by (44), gives (42).

Step 3. We now apply (42) to $f_1 = f_{A,C;\gamma_0}$ and $f_2 = f_{A,C;-\gamma_0}$, and where the characteristic functions are determined by (41) and $(\mathscr{F}_4 f_{A,C})$. We have already checked (43) for f_j. As for (44), 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,

$$\big(\mathscr{F}_4f_1\big)(u,v) - \big(\mathscr{F}_4f_1\big)(u,v) = 2\,\gamma_0\exp(3iu_1)\,Q\big(-i(u,v)\big)\,\big(\mathscr{F}_4G_1\big)(u,v).$$

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

$$\begin{aligned}
\partial_{u_1}\big(\mathscr{F}_4f_1\big)(u,v) - \big(\mathscr{F}_4f_1\big)(u,v) =&2\,\gamma_0\,Q\big(-i(u,v)\big)\,\big(3i\exp(3iu_1)\big(\mathscr{F}_4G_1\big)(u,v) + \exp(3iu_1)\,\partial_{u_1}\big(\mathscr{F}_4G_1\big)(u,v)\big)\\
&+2\,i\,v_0\,\gamma_0\exp(3iu_1)\,\big(\mathscr{F}_4G_1\big)(u,v),\qquad (u,v)\in\mathbb{R}^4.
\end{aligned}$$

Since

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

applying (42) yields

$$f_{B,1}(b_0,b_1) - f_{B,2}(b_0,b_1) = \frac{-2i\gamma_0}{(2\pi)^2} \int_{\mathbb{R}^2} v_0 \exp(-iv_0b_0) \exp(-3iv_1b_1) \left(\mathscr{F}_4 G_1\right)(-v_0b_1, -v_1b_1, v)\, dv. \tag{46}$$

Step 4. In the final step, we show that

$$\begin{aligned} &\int_{\mathbb{R}} b_0 \int_{\mathbb{R}} \left(f_{B,1}(b_0,b_1) - f_{B,2}(b_0,b_1)\right) db_1 db_0 \\ =&2\gamma_0 \int_{\mathbb{R}^4} b_0\, \eta(a_0,a_1,b_0,b_1)\, da\, db \neq 0. \end{aligned}$$

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 (46) 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

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

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

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

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

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

where

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

We obtain

$$\begin{aligned} \int_{\mathbb{R}} b_0 \left(\int_{\mathbb{R}} D(b_0,b_1)\, db_1\right) db_0 = \int_{\mathbb{R}} b_0 F(b_0)\, db_0 &= (-i) \frac{\mathrm{d}}{\mathrm{d}v_0} \left(\mathscr{F}F\right)(0) \\ &= 2\pi(-i) \frac{\mathrm{d}}{\mathrm{d}v_0} \tilde{F}(0) = 2\pi(-i) \int_{\mathbb{R}} E(b_1)\, db_1 \neq 0, \end{aligned}$$

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

Theorem 11. *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 11. 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

$$\int_{\mathbb{R}} b_2 \int_{\mathbb{R}^2}\int_1^\infty a_1\big(f_1-f_2\big)(a_1,a_2,a_1b_1,b_2+b_1a_2)\,da_1da_2db_1db_2\neq 0. \tag{47}$$

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 (40). 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 (47). 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 10 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 Theorem 2. We use again the reduced form version of the general model,

$$\begin{aligned}
Y&=C_0+C_1Z+C_2W,\\
X&=A_0+A_1Z+A_2W.
\end{aligned}$$

where $C=(C_0,C_1,C_2)$, $C_0=B_0+B_1A_0$, $C_2=B_2+B_1A_2$ and $C_1=B_1A_1$. 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 10, 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

$$\begin{aligned}
E\Big(\exp\big(it_1X+it_2Y\big)\big|Z=z,W=w\Big)&=E\Big(\exp\big(i(t_1zA_1+t_1wA_2+t_2zC_1+t_2wC_2)\big)\Big)\\
&=\psi_{A_1,A_2,C_1,C_2}(t_1z,t_1w,t_2z,t_2w),
\end{aligned}$$

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

$$\mathscr{S}'=\big\{(t_1z,t_1w,t_2z,t_2w)\in\mathbb{R}^4:\ t_1,t_2\in\mathbb{R},\ (z,w)\in\operatorname{supp}(Z,W)\big\},$$

and contains all the information from the joint distribution of the observations, in the sense analogous to Lemma 10. For the polynomial $Q(u_1,u_2,v_1,v_2)=u_1v_2-u_2v_1$ 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, and non-identification prevails.

∎

A.2. Proofs for Section 3

The following smoothness assumption is needed in the second part of Theorem 4 to justify interchanging the orders of integration.

Assumption 10.

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

Proof of Theorem 4. (i) We provide the missing details for the proof of the first part of Theorem 4. By simple change of variables, for fixed z, $f_{A_0+A_1z,A_1,B}(x,a_1,b) = f_{A_0,A_1,B}(x-a_1z,a_1,b)$, so that

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

and therefore

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

Moreover, by exogeneity of Z, $f_{A_0+A_1z}(x) = f_{X|Z}(x|z)$, which then implies (11). Next, we justify the first equality in (12). Note that we may only integrate z in the first line over the support of $Z|X=x$, which equals the support of $A_x = (x-A_0)/A_1$ since Z itself has full support. By change of variables,

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

so that it suffices to integrate out over this support in the second line as well.

The computation in (12) and Assumption 3 then lead to (13). Finally, by Fourier inversion,

$$\begin{aligned} f_B(b) &= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \exp\big(-i(b_0u_0+b_1u_1)\big)\, \big(\mathscr{F}_2 f_B\big)(u_0,u_1) du_0\, du_1 \\ &= \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} |t| \exp\big(-it\,(b_0+b_1x)\big)\, \big(\mathscr{F}_2 f_B\big)\,(t,tx) dt\, dx \end{aligned}$$

which together with (13) and the definition of T this implies the reconstruction formula (15).

The expression (16) for $E|A_1|^{-1}$ is obtained by setting $t=0$ in (13).

(ii) By the above, Assumption 10 is equivalent to

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

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

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

Using

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

yields (15). ∎

Proof of Theorem 5. 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,

$$\begin{aligned}\left(\mathscr{F}_{2+S}f_B\right)(t,tx,tw)\,E\left|A_{1,1}\right|^{-1} &= E\left(\exp\left(it(B_0+B_1x+B_2'w)\right)\right)E\left|A_{1,1}\right|^{-1}\\ &= \int_{\mathbb{R}^L}\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)\,f_{X|Z,W}(x|z,w)\,f_{Z_{-1}}(z_{-1})\,dz.\end{aligned} \tag{48}$$

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

$$\begin{aligned}\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w) =& E\left(\exp\left(it(B_0+B_1X+B_2'W)\right)|X=x,Z=z,W=w\right)\\ =& E\left(\exp\left(it(B_0+B_1x+B_2'w)\right)|A_0+A_1'z+A_2'w=x\right).\end{aligned} \tag{49}$$

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

$$\begin{aligned}&\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)\,f_{X|Z,W}(x|z,w)\\ &\quad= \int_{\mathbb{R}^{2+2S+L}}\exp\left(it(b_0+b_1x+b_2'w)\right)f_{B,A}(b,x-a_1'z-a_2'w,a_1,a_2)\,da_1\,da_2\,db.\end{aligned} \tag{50}$$

Integrating out z_1 gives

$$\begin{aligned}&\int_{\mathbb{R}}\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)\,f_{X|Z,W}(x|z,w)\,dz_1\\ &\quad= \int_{\mathbb{R}^{3+2S+L}}\exp\left(it(b_0+b_1x+b_2'w)\right)f_{B,A}(b,x-a_1'z-a_2'w,a_1,a_2)\,dz_1\,da_1\,da_2\,db\\ &\quad= \int_{\mathbb{R}^{3+2S+L}}|a_{1,1}|^{-1}\exp\left(it(b_0+b_1x+b_2'w)\right)f_{B,A}(b,a)\,da\,db\\ &\quad= E\left(\exp\left(it(B_0+B_1x+B_2w)\right)\frac{1}{|A_{1,1}|}\right) = E\left(\exp\left(it(B_0+B_1x+B_2w)\right)\right)E\left|A_{1,1}\right|^{-1}.\end{aligned} \tag{51}$$

using a change of variables and Assumption 5 in the last step. Averaging over the values of Z_{-1} then gives (48), and applying the operator T_{S+1} we obtain (17). Finally, taking $t=0$ in (51), we see that for any x,z_{-1},w, (18) holds true. ∎

Proof of Theorem 6. 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 (50) we obtain that

$$\begin{aligned}&\mathscr{F}_1\left(f_{Y|X,Z,W}\right)(t|x,z,w)\,f_{X|Z,W}(x|z,w)\\ &\quad= \int_{\mathbb{R}^{3+2S}}\exp\left(it(b_0+b_1x+b_2'w)\right)|a_1|^{-1}f_{A_x,A_1,A_2,B}(z,a_1,a_2,b)\,da_1\,da_2\,db.\end{aligned} \tag{52}$$

Under the support assumption (19), it suffices to integrate out z over the support of the conditional distribution of $Z|W=w$ to obtain (20). ∎

Proof of Theorem 7. First, we require the following lemma, which does not involve the model itself but only a set of random coefficients.

Lemma 12. *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*

$$f_B(b_0,b_1)=\frac{1}{(2\pi)^2}\int_{\mathbb{R}}\int_{\mathbb{R}}\exp(-itb_0)\psi_{A,C}(-tb_1,-tzb_1,t,tz)\,|t|\,dt\,dz\left(E|A_1|^{-1}\right)^{-1}. \tag{53}$$

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 4, *(ii)*, are satisfied, and we obtain (15). Using the equality (5) immediately gives (53). ∎

The following lemma is based on analytic continuation.

Lemma 13. *Under Assumption 7, for any fixed t and b_1 the function*

$$\Phi: z\mapsto \psi_{A,C}(-tb_1,-tb_1z,t,tz),$$

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

Proof of Lemma 13: 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

$$\Phi_0: z\mapsto \psi_{A,C}(-tb_1,-tb_1z,t,tz),\qquad \Phi_1: z\mapsto \psi_{\tilde{A},\tilde{C}}(-tb_1,-tb_1z,t,tz)$$

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

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

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

$$\Phi^{(k+1)}(z)=\sum_{l=0}^{k+1}\binom{k+1}{l}(-tb_1)^l t^{k+1-l}\frac{\partial^{k+1}\psi}{(\partial a_1)^l(\partial c_1)^{k+1-l}}(-tb_1,-tb_1z,t,tz).$$

Since

$$\left|\frac{\partial^{k+1}\psi}{(\partial a_1)^l(\partial c_1)^{k+1-l}}(-tb_1,-tb_1z,t,tz)\right|\le E|A_1|^l|C_1|^{k+1-l}+E|\tilde{A}_1|^l|\tilde{C}_1|^{k+1-l},$$

it follows by binomial expansion that

$$\begin{aligned}|\Phi^{(k+1)}(z)|&\le\sum_{l=0}^{k+1}\binom{k+1}{l}|tb_1|^l|t|^{k+1-l}\left(E|A_1|^l|C_1|^{k+1-l}++E|\tilde{A}_1|^l|\tilde{C}_1|^{k+1-l}\right)\\&=|t|^{k+1}\left(E\left(|b_1A_1|+|C_1|\right)^{k+1}+E\left(|b_1\tilde{A}_1|+|\tilde{C}_1|\right)^{k+1}\right)\\&\le 2^{k+1}|tb_1|^{k+1}\left(E|A_1|^{k+1}+E|\tilde{A}_1|^{k+1}\right)+2^{k+1}|t|^{k+1}\left(E|C_1|^{k+1}+E|\tilde{C}_1|^{k+1}\right).\end{aligned}$$

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 7 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 13, 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 (53),

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

we identify f_B. ∎

Proof of Theorem 8. Given $g \geq 0$ and $p > 1$ such that $EA_g^p < \infty$, we have from the Hölder-inequality that

$$G(t; x, w, g) = E\left(1_{Y_{x,w} \leq t} A_g\right) \leq \left(E\left(A_1^p\right)\right)^{1/p} \left(F_{Y_{x,w}}(t)\right)^{(p-1)/p}.$$

Solving for $F_{Y_{x,w}}(t)$ gives the lower bound.
For the upper bound, given $g > 0$ and $p > 1$ with $EA_g^{-p} < \infty$. 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

$$\begin{aligned} F_{Y_{x,w}}(t) &= E\left(\left(1_{Y_{x,w} \leq t} A_g\right)^{1/r} A_g^{-1/r}\right) \\ &\leq \left(G(t; x, w, g)\right)^{1/r} E\left(A_g^{-s/r}\right)^{1/s} = \left(G(t; x, w, g)\right)^{p/(p+1)} E\left(A_g^{-p}\right)^{1/(p+1)}. \end{aligned}$$

∎

B. Semiparametric estimation: Technical assumptions, results and proofs

Lemma 14. *Let $\Gamma(x)$ denote the gamma function. For any $\kappa \geq 2$ and $n \geq 2$, we have under Assumption 8 that*

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

The proof is deferred to the technical supplement, Hoderlein et al. (2016a), Section D.

Assumption 11.

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

Lemma 15. *Suppose that f_Z is continuous on $[-1, 1]$. Given Assumption 8 we have for any uniformly Lipschitz continuous function $f : [-1, 1] \to \mathbb{R}$ that*

$$E\Big| \sum_{j=1}^{n-1} f(Z_{(j)}) \cdot (Z_{(j+1)} - Z_{(j)}) - \int_{-1}^{1} f(z) dz \Big|^2 = \mathscr{O}(n^{-2}), \tag{54}$$

$$\lim_{n \to \infty} E\Big| n \sum_{j=1}^{n-1} f(Z_{(j)}) \cdot (Z_{(j+1)} - Z_{(j)})^2 - 2\int_{-1}^{1} f(z)/f_Z(z)\, dz \Big|^2 = 0. \tag{55}$$

Proof of Lemma 15: In order to prove the claim (**??**) we consider that

$$\begin{aligned}
&E\Big|\sum_{j=1}^{n-1} f(Z_{(j)})\cdot(Z_{(j+1)}-Z_{(j)}) - \int_{-1}^{1} f(z)dz\Big|^2 \\
&\leq 2E\Big(\sum_{j=1}^{n-1}\int_{Z_{(j)}}^{Z_{(j+1)}} |f(Z_j)-f(z)|dz\Big)^2 + 4\|f\|_\infty \cdot E|Z_{(n)}-1|^2 + 4\|f\|_\infty \cdot E|Z_{(1)}+1|^2 \\
&\leq 2C_L E\sum_{j=1}^{n-1}(Z_{(j+1)}-Z_{(j)})^2 + 4\|f\|_\infty \cdot E|Z_{(n)}-1|^2 + 4\|f\|_\infty \cdot E|Z_{(1)}+1|^2,
\end{aligned}$$

where $C_L := \sup\{|f(x)-f(y)|/|x-y| : x \neq y \in [-1,1]\}$ denotes the Lipschitz constant of the function f. Then, applying Lemma 14 completes the proof of (**??**).

Now we focus on (54). We use that

$$\sum_{j=1}^{n-1} f(Z_{(j)})(Z_{(j+1)}-Z_{(j)})^2 = \sum_{j=1}^{n} f(Z_j)(Z_j^*-Z_j)^2,$$

holds true almost surely where

$$Z_j^* := \begin{cases} \min\{Z_k : Z_k > Z_j\}, & \text{if } Z_j < Z_{(n)}, \\ Z_{(n)}, & \text{otherwise,} \end{cases}$$

as in the proof of Lemma 14. We define

$$T_n := n\sum_{j=1}^{n-1} f(Z_{(j)})\cdot(Z_{(j+1)}-Z_{(j)})^2.$$

First we study the expectation of T_n.

$$ET_n = E\Big\{n\sum_{j=1}^{n-1} f(Z_{(j)})\cdot(Z_{(j+1)}-Z_{(j)})^2\Big\} = n\sum_{j=1}^{n} E\,f(Z_j)\cdot E\big\{(Z_j^*-Z_j)^2 \mid Z_j\big\}. \tag{56}$$

Since

$$\begin{aligned}
E\big\{(Z_j^*-Z_j)^2 \mid Z_j\big\} &= \int_{s=0}^{\infty} P\big[Z_j^* > Z_j + \sqrt{s} \mid Z_j\big]ds \\
&= \int_{s=0}^{(1-Z_j)^2}\Big(1-\int_{Z_j}^{Z_j+\sqrt{s}} f_Z(u)du\Big)^{n-1}ds - \int_{s=0}^{(1-Z_j)^2}\Big(1-\int_{Z_j}^{\infty} f_Z(u)du\Big)^{n-1}ds,
\end{aligned}$$

holds true almost surely, (56) equals

$$n^2\int f(z)f_Z(z)\int_{s=0}^{(1-z)^2}\Big(1-\int_z^{z+\sqrt{s}} f_Z(u)du\Big)^{n-1}dsdz - n^2\int f(z)f_Z(z)(1-z)^2\Big(1-\int_z^{\infty} f_Z(u)du\Big)^{n-1}dz. \tag{57}$$

The second term in (57) obeys the upper bound

$$\begin{aligned} n^2\|f\|_\infty\|f_Z\|_\infty \int_{-1}^{1}(1-z)^2\exp\big(-(n-1)c_Z(1-z)\big)dz &\le n^2\|f\|_\infty\|f_Z\|_\infty\int_0^\infty v^2\exp\big(-(n-1)c_Z v\big)dv \\ &= n^2(n-1)^{-3}\|f\|_\infty\|f_Z\|_\infty\int_0^\infty v^2\exp(-c_Z v)dv \\ &= \mathscr{O}(1/n)\,, \end{aligned}$$

so that this term is asymptotically negligible. By the integral substitution $w = n^2 s$ the first term in (57) equals

$$\int\int_{w=0}^{\infty} f(z)f_Z(z)\,1_{[0,n^2(1-z)^2]}(w)\left(1-\int_z^{z+\sqrt{w}/n} f_Z(u)du\right)^{n-1}dw\,dz\,.$$

As

$$\left|f(z)f_Z(z)\,1_{[0,n^2(1-z)^2]}(w)\left(1-\int_z^{z+\sqrt{w}/n} f_Z(u)du\right)^{n-1}\right| \le \|f\|_\infty f_Z(z)\exp\big(-c_Z\sqrt{w}/2\big)\,,$$

for all $z\in\mathbb{R}$, $w\ge 0$ and $n\ge 2$ it follows by dominated convergence that the term (57), and hence ET_n, tends to

$$\int f(z)f_Z(z)\int_{w=0}^{\infty}\exp\big(-f_Z(z)\sqrt{w}\big)dw\,dz = 2\int_{-1}^{1} f(z)/f_Z(z)dz, \tag{58}$$

as $n\to\infty$.

Now we consider the second moment of T_n. We have that

$$ET_n^2 = n^2\sum_{j,j'=1}^{n} Ef(Z_j)f(Z_{j'})E\big\{(Z_j^*-Z_j)^2(Z_{j'}^*-Z_{j'})^2\mid Z_j,Z_{j'}\big\}\,. \tag{59}$$

The partial sum of the diagonal terms (for $j=j'$) in (59) equals

$$n^2\sum_{j=1}^{n} Ef^2(Z_j)E\big\{(Z_j^*-Z_j)^4\mid Z_j\big\} \le n^2\|f\|_\infty^2 E\sum_{j=1}^{n-1}(Z_{(j+1)}-Z_{(j)})^4 = \mathscr{O}(1/n)\,, \tag{60}$$

by Lemma 14. For all $j\neq j'$ we consider that

$$\begin{aligned} P\big[Z_{j'}^*>Z_{j'}+t, Z_j^*>Z_j+s\mid Z_j,Z_{j'}\big] = 1_{[0,Z_{j'}-Z_j]}(s)\cdot\Big\{&\left(1-\int_{Z_j}^{Z_j+s}f_Z(u)du-\int_{Z_{j'}}^{Z_{j'}+t}f_Z(u)du\right)^{n-1} \\ &-\left(1-\int_{Z_j}^{Z_j+s}f_Z(u)du-\int_{Z_{j'}}^{\infty}f_Z(u)du\right)^{n-1}\Big\}\,, \end{aligned}$$

for all $s,t>0$ on the event $\{Z_j<Z_{j'}\}$. Otherwise, if $Z_j>Z_{j'}$, then

$$\begin{aligned} P\big[Z_{j'}^*>Z_{j'}+t, Z_j^*>Z_j+s\mid Z_j,Z_{j'}\big] = 1_{[0,Z_j-Z_{j'}]}(t)\cdot\Big\{&\left(1-\int_{Z_j}^{Z_j+s}f_Z(u)du-\int_{Z_{j'}}^{Z_{j'}+t}f_Z(u)du\right)^{n-1} \\ &-\left(1-\int_{Z_{j'}}^{Z_{j'}+t}f_Z(u)du-\int_{Z_j}^{\infty}f_Z(u)du\right)^{n-1}\Big\}\,. \end{aligned}$$

Taking the second order partial derivative with respect to s and t, the conditional joint density of $Z_j^*-Z_j$

and $Z^*_{j'} - Z_{j'}$ given Z_j and $Z_{j'}$ turns out to be

$$f_{(Z^*_j - Z_j, Z^*_{j'} - Z_{j'})|Z_j, Z_{j'}}(s,t) = \big[1_{[0, Z_{j'} - Z_j]}(s) \cdot 1\{Z_j < Z_{j'}\} + 1_{[0, Z_j - Z_{j'}]}(t) \cdot 1\{Z_j > Z_{j'}\}\big] \cdot (n-1)(n-2)\Big(1 - \int_{Z_j}^{Z_j+s} f_Z(u)du - \int_{Z_{j'}}^{Z_{j'}+t} f_Z(u)du\Big)^{n-3} f_Z(Z_j+s) f_Z(Z_{j'}+t),$$

for $n \geq 4$. Therefore, if $Z_j < Z_{j'}$, we have

$$\begin{aligned} &E\big\{(Z^*_j - Z_j)^2 (Z^*_{j'} - Z_{j'})^2 \mid Z_j, Z_{j'}\big\} = \iint s^2 t^2 f_{(Z^*_j - Z_j, Z^*_{j'} - Z_{j'})|Z_j, Z_{j'}}(s,t) ds\, dt \\ &= \frac{(n-1)(n-2)}{(n-3)^6} \int_{t=0}^{\infty} \int_{s=0}^{\infty} 1_{[0,(n-3)(1-Z_{j'})]}(t) 1_{[0,(n-3)(Z_{j'} - Z_j)]}(s) \\ &\cdot s^2 t^2 \cdot \Big(1 - \int_{Z_j}^{Z_j + s/(n-3)} f_Z(u)du - \int_{Z_{j'}}^{Z_{j'} + t/(n-3)} f_Z(u)du\Big)^{n-3} f_Z(Z_j + s/(n-3)) f_Z(Z_{j'} + t/(n-3))\, ds\, dt\,. \end{aligned} \tag{61}$$

The function to be integrated in (61) is bounded by

$$\|f_Z\|_\infty^2 s^2 t^2 \exp\big(-c_Z[s+t]\big), \qquad s,t > 0\,. \tag{62}$$

By dominated convergence the integral in (61) converges to

$$f_Z(Z_j) f_Z(Z_{j'}) \iint_{s,t>0} s^2 t^2 \exp\big(-f_Z(Z_j)s\big) \exp\big(-f_Z(Z_j)t\big) ds\, dt = 4/\{f_Z^2(Z_j) f_Z^2(Z_{j'})\},$$

as $n \to \infty$. Otherwise, if $Z_j > Z_{j'}$, the same convergence and the same bound as in (62) occur. We have

$$\begin{aligned} &n^2 \sum_{j \neq j'=1}^{n} E\, f(Z_j) f(Z_{j'}) E\big\{(Z^*_j - Z_j)^2 (Z^*_{j'} - Z_{j'})^2 \mid Z_j, Z_{j'}\big\} \\ &= \frac{n^3 (n-1)^2 (n-2)}{(n-3)^6} \iint f(z) f(z') f_Z(z) f_Z(z') E\big\{(Z^*_j - Z_j)^2 (Z^*_{j'} - Z_{j'})^2 \mid Z_j = z, Z_{j'} = z'\big\} dz\, dz'\,. \end{aligned} \tag{63}$$

The term (62) provides an upper bound for the function to to integrated in (63) as well. Again, by dominated convergence, we derive that (63) tends to

$$\iint f(z) f(z') f_Z(z) f_Z(z') \cdot 4/\{f_Z^2(z) f_Z^2(z')\} dz\, dz' = 4\Big(\int_{-1}^{1} f(z)/f_Z(z) dz\Big)^2, \tag{64}$$

as $n \to \infty$. Combining this with (59) and (60) it follows that ET_n^2 converges to (64). Together with (58) this implies that the variance of T_n converges to 0. Considering the convergence (58) again we have finally verified (54). □

Lemma 16. *Suppose that f_Z is continuous on $[-1,1]$ and that Assumptions 8 and 11 hold true. For a bounded function $f : \mathbb{R}^2 \times [-1,1] \to \mathbb{R}^d$ we define the sequence of random vectors*

$$\zeta_n = \sqrt{n}\Big(\sum_{j=1}^{n-1} f(Y_{(j)}, X_{(j)}, Z_{(j)}) \cdot (Z_{(j+1)} - Z_{(j)}) - \int_{-1}^{1} \mu(z)\, dz\Big),$$

where $\mu(z) := E\big[f(Y_1,X_1,Z_1) \mid Z_1 = z\big]$. *Then* (ζ_n) *is asymptotically normally distributed*

$$\zeta_n \xrightarrow{d} N\Big(0, 2\int_{-1}^{1} \sigma^2(z)/f_Z(z)\,dz\Big),$$
$$\sigma^2(z) = \big[cov\{f_j(Y_1,X_1,Z_1), f_k(Y_1,X_1,Z_1) \mid Z_1 = z\}\big]_{j,k=1,\ldots,d}.$$

Proof of Lemma 16: The conditional characteristic function of ζ_n given σ_Z turns out to be

$$\Psi_{\zeta_n|\sigma_Z}(t) = \exp\Big(it'\sqrt{n}\Big[\sum_{l=1}^{n-1}\mu(Z_{(l)})\,(Z_{(l+1)} - Z_{(l)}) - \int_{-1}^{1}\mu(z)dz\Big]\Big)$$
$$\cdot \prod_{l=1}^{n-1} E\Big(\exp\Big\{it'\sqrt{n}\big(f(Y_{(l)},X_{(l)},Z_{(l)}) - \mu(Z_{(l)})\big)\,(Z_{(l+1)} - Z_{(l)})\Big\}\Big|\sigma_Z\Big), \tag{65}$$

for any $t \in \mathbb{R}^d$, using the conditional independence of the $(Y_{(l)},X_{(l)},Z_{(l)})$, $l = 1,\ldots,n-1$, given σ_Z. By Taylor approximation we deduce that

$$\Big|\prod_{l=1}^{n-1} E\Big(\exp\Big\{it'\sqrt{n}\big(f(Y_{(l)},X_{(l)},Z_{(l)}) - \mu(Z_{(l)})\big)(Z_{(l+1)} - Z_{(l)})\Big\}\Big|\sigma_Z\Big)$$
$$- \exp\Big(-\frac{1}{2}t'n\sum_{l=1}^{n-1}\sigma^2(Z_{(l)})(Z_{(l+1)} - Z_{(l)})^2 t\Big)\Big|$$
$$\le \frac{2}{3}n^{3/2}|t|^3\|f\|_\infty^3\sum_{l=1}^{n-1}(Z_{(l+1)} - Z_{(l)})^3 + n^2|t|^4\|f\|_\infty^4\sum_{l=1}^{n-1}(Z_{(l+1)} - Z_{(l)})^4,$$

so that, by Lemma 14 and (65), we have that

$$\Big|\Psi_{\zeta_n}(t) - E\exp\Big(it'\sqrt{n}\Big[\sum_{l=1}^{n-1}\mu(Z_{(l)})\,(Z_{(l+1)} - Z_{(l)}) - \int_{-1}^{1}\mu(z)dz\Big] - \frac{1}{2}t'n\sum_{l=1}^{n-1}\sigma^2(Z_{(l)})(Z_{(l+1)} - Z_{(l)})^2 t\Big)\Big|$$
$$= \mathscr{O}\big(n^{-1/2}\big). \tag{66}$$

Assumption 11 guarantees that the functions μ and σ^2 are uniformly Lipschitz continuous on the domain $[-1,1]$. Then applying Lemma 15 to each component of μ and σ^2 in (66) completes the proof as pointwise convergence of the characteristic function implies convergence in distribution. □

From Lemma 16 we immediately obtain the following result.

Proposition 17. *Given Assumptions 1, 8 and 11, for an interval* $I \subseteq suppX$ *for which (25) is satisfied, we have for the estimator* $\hat{a}_{I,n}$ *of the scaling constant in (28) that*

$$\sqrt{n}\,\big(\hat{a}_{I,n} - E|A_1|^{-1}\big) \xrightarrow{d} N\Big(0, 2\int_{-1}^{1}\sigma^2(z)/f_Z(z)\,dz\Big),$$
$$\sigma^2(z) = E\phi^2(A_0 + A_1 z) - \big[E\phi(A_0 + A_1 z)\big]^2$$

We present the proof of Theorem 9 in two steps, and start with the rate of convergence (33)

Proof of Theorem 9, rate of convergence. For the proof let $s_c = E|A_1|^{-1} > 0$. We start by showing the consistency of $\hat{\theta}_n$, by checking the conditions in van der Vaart (1998, Theorem 5.7). Since Θ is assumed to be compact and

$$\theta \mapsto \|\Phi(\theta,\cdot) - \Phi(\theta_0,\cdot)\|_{v;q}$$

is continuous by dominated convergence, Assumption 9 (first part) implies the second condition in van der Vaart (1998, Theorem 5.7). Let

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

for which $0 < C_\Theta < \infty$. Then by the triangle inequality,

$$\sup_{\theta\in\Theta} \left| \|\hat\Phi_n(\cdot) - \hat a_{n;I}\Phi(\theta,\cdot)\|_{v;q} - \|\hat\Phi_n(\cdot) - s_c\Phi(\theta,\cdot)\|_{v;q} \right| \le \left|\hat a_{n;I} - s_c\right| C_\Theta,$$

which is $o_P(1)$ from Proposition 17. Moreover, by the triangle inequality,

$$\begin{aligned} &\sup_{\theta\in\Theta} \left| \|\hat\Phi_n(\cdot) - s_c\Phi(\theta,\cdot)\|_{v;q} - \|s_c\Phi(\theta_0,\cdot) - s_c\Phi(\theta,\cdot)\|_{v;q} \right| \\ \le &\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q} + \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_0,\cdot)\|_{v;q} \end{aligned} \tag{67}$$

Below we show for integer $\kappa \ge 2$ that

$$E\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q}^\kappa = \mathscr{O}(n^{-\kappa/2}), \tag{68}$$

and that

$$\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_0,\cdot)\|_{v;q} = \mathscr{O}(n^{-1}), \tag{69}$$

which together with (67) imply the condition on uniform convergence in van der Vaart (1998, Theorem 5.7), and hence the consistency.

To obtain the rate of convergence, given $\varepsilon > 0$ we find $n_0, M \in \mathbb{N}$ such that

$$P\left[\sqrt{n}\,\|\hat\theta_n - \theta_0\| > 2^M\right] \le 2\varepsilon,$$

for all $n \ge n_0$. From Proposition 17, choose $\tilde M > 0$ so large that

$$\begin{aligned} &P\left[\sqrt{n}\,\|\hat\theta_n - \theta_0\| > 2^M\right] \\ \le &P\left[\|\sqrt{n}\,\|\hat\theta_n - \theta_0\| > 2^M,\ \left|\hat a_{I;n} - s_c\right| \le \tilde M/\sqrt{n}\right] + P\left[\left|\hat a_{I;n} - s_c\right| > \tilde M/\sqrt{n}\right] \\ \le &P\left[\sqrt{n}\,\|\hat\theta_n - \theta_0\| > 2^M,\ \left|\hat a_{I;n} - s_c\right| \le \tilde M/\sqrt{n}\right] + \varepsilon \end{aligned}$$

for all n. Let

$$S_{j,n} = \left\{\theta\in\Theta:\ 2^{j-1} < \sqrt{n}\,\|\hat\theta_n - \theta_0\| \le 2^j\right\}.$$

If $M, n \in \mathbb{N}$ are such that $\varepsilon_0 > 2^M/\sqrt{n}$, where ε_0 is as in Assumption 9 (second part), then

$$\begin{aligned} &P\left[\sqrt{n}\,\|\hat\theta_n - \theta_0\| > 2^M,\ \left|\hat a_{I;n} - s_c\right| \le \tilde M/\sqrt{n}\right] \\ \le &P\left[\|\hat\theta_n - \theta_0\| > \varepsilon_0\right] + \sum_{j\ge M+1, 2^{j-1}\le\sqrt{n}\varepsilon_0} P\left|\hat\theta_n \in S_{j,n},\ \left|\hat a_{I;n} - s_c\right| \le \tilde M/\sqrt{n}\right], \end{aligned}$$

where the first term is $\le \varepsilon/2$ for large n by consistency. For the second,

$$\begin{aligned} &P\left|\hat\theta_n \in S_{j,n},\ \left|\hat a_{I;n} - s_c\right| \le \tilde M/\sqrt{n}\right] \\ \le &P\left[\inf_{\theta\in S_{j,n}} \|\hat\Phi_n(\cdot) - \hat a_{I;n}\Phi(\theta,\cdot)\|_{v;q} \le \|\hat\Phi_n(\cdot) - \hat a_{I;n}\Phi(\theta_0,\cdot)\|_{v;q},\ \left|\hat a_{I;n} - s_c\right| \le \tilde M/\sqrt{n}\right] \\ \le &P\left[\inf_{\theta\in S_{j,n}} \|\hat\Phi_n(\cdot) - s_c\Phi(\theta,\cdot)\|_{v;q} \le \|\hat\Phi_n(\cdot) - s_c\Phi(\theta_0,\cdot)\|_{v;q} + 2C_\Theta\tilde M/\sqrt{n}\right] \\ \le &P\left[\inf_{\theta\in S_{j,n}} \max\left(\|E\hat\Phi_n(\cdot) - s_c\Phi(\theta,\cdot)\|_{v;q} - \|E\hat\Phi_n(\cdot) - s_c\Phi(\theta_0,\cdot)\|_{v;q} - 2C_\Theta\tilde M/\sqrt{n}, 0\right) \le 2\,\|\hat\Phi_n(\cdot) - E\hat\Phi_n(\cdot)\|_{v;q},\right] \end{aligned}$$

by the triangle inequality, since

$$\|\hat\Phi_n(\cdot) - s_c\Phi(\theta,\cdot)\|_{v;q} \le \|\hat\Phi_n(\cdot) - s_c\Phi(\theta_0,\cdot)\|_{v;q} + 2C_\Theta\tilde M/\sqrt{n}$$

implies

$$\begin{aligned}&\max\Big(\|E\hat{\Phi}_n(\cdot)-s_c\,\Phi(\theta,\cdot)\|_{v;q}-\|E\hat{\Phi}_n(\cdot)-s_c\,\Phi(\theta_0,\cdot)\|_{v;q}-2C_\Theta\tilde{M}/\sqrt{n},0\Big)\\ \le&2\,\|\hat{\Phi}_n(\cdot)-E\hat{\Phi}_n(\cdot)\|_{v;q}.\end{aligned}$$

Below, using Assumption 9 (second part) we show for $\|\theta-\theta_0\|\le\varepsilon_0$ that

$$\|E\hat{\Phi}_n(\cdot)-s_c\,\Phi(\theta,\cdot)\|_{v;q}-\|E\hat{\Phi}_n(\cdot)-s_c\,\Phi(\theta_0,\cdot)\|_{v;q}\ge c_\Theta\, s_c\,\|\theta_0-\theta\|-\mathcal{O}(1/n), \tag{70}$$

where the remainder term is uniform in θ. Therefore

$$\begin{aligned}&P\big[\big|\hat{\theta}_n\in S_{j,n},\ \big|\hat{a}_{I;n}-s_c\big|\le\tilde{M}/\sqrt{n}\big]\\ \le&P\Big[\frac{1}{2}\big(c_\Theta\, s_c\,2^{j-1}-2C_\Theta\tilde{M}-\mathcal{O}(1/\sqrt{n})\big)/\sqrt{n}\le\|\hat{\Phi}_n(\cdot)-E\hat{\Phi}_n(\cdot)\|_{v;q}\Big],\end{aligned}$$

where the lower bound is strictly positive for large j. Hence, for large n,

$$\begin{aligned}&P\big[\sqrt{n}\,\|\hat{\theta}_n-\theta_0\|>2^M,\ \big|\hat{a}_{I;n}-s_c\big|\le\tilde{M}/\sqrt{n}\big]\\ \le&\,\varepsilon/2+\sum_{j\ge M+1}P\Big[\frac{1}{2}\big(c_\Theta\, s_c\,2^{j-1}-2C_\Theta\tilde{M}-\mathcal{O}(1/\sqrt{n})\big)/\sqrt{n}\le\|\hat{\Phi}_n(\cdot)-E\hat{\Phi}_n(\cdot)\|_{v;q}\Big]\\ \le&\,\varepsilon/2+4\sum_{j\ge M+1}\frac{nE\|\hat{\Phi}_n(\cdot)-E\hat{\Phi}_n(\cdot)\|_{v;q}^2}{\frac{1}{2}\big(c_\Theta\, s_c\,2^{j-1}-2C_\Theta\tilde{M}-\mathcal{O}(1/\sqrt{n})\big)}\\ =&\,\varepsilon/2+\mathcal{O}(1)\sum_{j\ge M+1}\frac{1}{\frac{1}{2}\big(c_\Theta\, s_c\,2^{j-1}-2C_\Theta\tilde{M}-\mathcal{O}(1/\sqrt{n})\big)}\end{aligned}$$

by (68), and the last sum is arbitrarily small for large M, uniformly for large n.

It remains to prove (68) and (70). Using the simple inequality $(a+b)^\kappa\le 2^\kappa(a^\kappa+b^\kappa)$, $a,b>0$, $\kappa\in\mathbb{N}$, we estimate

$$\begin{aligned}&E\big\|\hat{\Phi}_n(\cdot)-E\hat{\Phi}_n(\cdot)\big\|_{v;q}^\kappa\le\frac{1}{q}\sum_{p=1}^q\int_{\mathbb{R}}\big|\hat{\Phi}_n(t,I_p)-E\hat{\Phi}_n(t,I_p)\big|^\kappa d\nu(t)\\ \le&\frac{2^\kappa}{q}\sum_{p=1}^q\int_{\mathbb{R}}\Big(\big|\hat{\Phi}_n(t,I_p)-E\big(\hat{\Phi}_n(t,I_p)|\sigma_Z\big)\big|^\kappa+\big|\mathbb{E}\big(\hat{\Phi}_n(t,I_p)|\sigma_Z\big)-E\hat{\Phi}_n(t,I_p)\big|^\kappa\Big)\,d\nu(t),\end{aligned} \tag{71}$$

We have that

$$\begin{aligned}E\big(\hat{\Phi}_n(t,I_p)\mid\sigma_Z\big)&=\sum_{j=1}^{n-1}E\Big(\exp(itY_{(j)})1_{I_p}(X_{(j)})\big|\sigma_Z\Big)\cdot\big(Z_{(j+1)}-Z_{(j)}\big)\\ &=\sum_{j=1}^{n-1}\big(Z_{(j+1)}-Z_{(j)}\big)\iint\exp(ity')1_{I_p}(x')\,f_{Y,X|Z}(y',x'|Z_{(j)})\,dy'\,dx'\\ &=\sum_{j=1}^{n-1}\int_{z=Z_{(j)}}^{Z_{(j+1)}}\iint\exp(ity')1_{I_p}(x')\,f_{Y,X|Z}(y',x'|Z_{(j)})\,dy'\,dx'\,dz.\end{aligned} \tag{72}$$

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

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

Therefore

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

again by Hölder's inequality. Taking the expectation on both sides of the above inequality and using Lemma 14 yields uniformly in t that

$$E\big|\hat\Phi_n(t,I_p)-E(\hat\Phi_n(t,I_p)\mid\sigma_Z)\big|^\kappa = \mathscr{O}(n^{-\kappa/2}). \tag{73}$$

To proceed, first observe that

$$\big|E\hat\Phi_n(t,I_p)-s_c\Phi(\theta_0,t,I_p)\big|^\kappa \le E\big|E(\hat\Phi_n(t,I_p)\mid\sigma_Z)-s_c\Phi(\theta_0,t,I_p)\big|^\kappa,\qquad \kappa\in\mathbb{N}. \tag{74}$$

Therefore

$$E\big|E(\hat\Phi_n(t,I_p)\mid\sigma_Z)-E\hat\Phi_n(t,I_p)\big|^\kappa \le 2^{2\kappa}E\big|E(\hat\Phi_n(t,I_p)\mid\sigma_Z)-s_c\Phi(\theta_0,t,I_p)\big|^\kappa. \tag{75}$$

To estimate the right side, note that

$$f_{Y,X|Z}(y,x|z) = \iint f_{A,B}(x-a_1z,a_1,y-c_1x,c_1)\,da_1\,dc_1\,,$$

so that, by Assumption 11, 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 (72) and (29) yields that

$$\big|E(\hat\Phi_n(t,I_p)\mid\sigma_Z)-a_1\Phi(\theta_0,t,I_p)\big| \le 2C_A\sum_{j=1}^{n-1}\big(Z_{(j+1)}-Z_{(j)}\big)^2+|Z_{(1)}+1|+|Z_{(n)}-1|, \tag{76}$$

holds true almost surely. Therefore, using Lemma 14, we get that

$$\begin{aligned}
&E\big|E(\hat\Phi_n(t,I_p)\mid\sigma_Z)-E\hat\Phi_n(t,I_p)\big|^\kappa\\
&\le 2^{2\kappa}E\Big(2C_A\sum_{j=1}^{n-1}\big(Z_{(j+1)}-Z_{(j)}\big)^2+|Z_{(1)}+1|+|Z_{(n)}-1|\Big)^\kappa\\
&\le 2C_A^\kappa 6^\kappa E\Big[\Big|\sum_{j=1}^{n-1}\big(Z_{(j+1)}-Z_{(j)}\big)^2\Big|^\kappa+E|Z_{(1)}+1|^\kappa+E|Z_{(n)}-1|^\kappa\\
&\le 2C_A^\kappa 6^\kappa(n-1)^{\kappa-1}E\sum_{j=1}^{n-1}\big(Z_{(j+1)}-Z_{(j)}\big)^{2\kappa}+E|Z_{(1)}+1|^\kappa+E|Z_{(n)}-1|^\kappa\\
&= \mathscr{O}(n^{-\kappa}).
\end{aligned} \tag{77}$$

Together with (73) and (71) this implies (68).

From (73), (76) and Lemma 14 we get that uniformly in t,

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

which directly yields (69). To obtain (70), for $\|\theta-\theta_0\| \le \varepsilon_0$ we estimate

$$\begin{aligned}
&\|E\hat{\Phi}_n(\cdot) - s_c\,\Phi(\theta,\cdot)\|_{\nu;q} - \|E\hat{\Phi}_n(\cdot) - s_c\,\Phi(\theta_0,\cdot)\|_{\nu;q} \\
\ge\, & s_c\,\|\Phi(\theta_0,\cdot) - \Phi(\theta,\cdot)\|_{\nu;q} - \mathscr{O}(1/n) \\
\ge\, & s_c\, c_\Theta^{1/2}\|\theta_0-\theta'\| - \mathscr{O}(1/n)\,,
\end{aligned}$$

by Assumption 9 (second part). This completes the proof of the theorem. ∎

Proof of Theorem 9, asymptotic normality. First we consider the term

$$\Delta_n := \Big\{\Big\langle \frac{\partial}{\partial\theta_j}\Phi(\theta_0,\cdot), \hat{\Phi}_n(\cdot) - \hat{a}_{I,n}\Phi(\theta_0,\cdot)\Big\rangle_{\nu;q}\Big\}_{j=1,\dots,d} = \sum_{l=1}^{n-1} \xi_l\cdot(Z_{(l+1)} - Z_{(l)})\,,$$

where $\langle\cdot,\cdot\rangle_{\nu;q}$ denotes the inner product

$$\langle\Phi_1,\Phi_2\rangle_{\nu;q} = \frac{1}{q}\sum_{j=1}^{q}\int_{\mathbb{R}} \Phi_1(t,I_j)\overline{\Phi_2(t,I_j)}d\nu(t)\,,$$

and

$$\xi_l := f(Y_{(l)},X_{(l)},Z_{(l)}) := \frac{1}{q}\sum_{k=1}^{q}\int_{\mathbb{R}} \nabla_\theta\Phi(\theta_0,t,I_k)\big\{\exp(-itY_{(l)})1_{I_k}(X_{(l)}) - \phi(X_{(l)})\overline{\Phi(\theta_0,t,I_k)}\big\}d\nu(t)\,.$$

Note that all ξ_l are $\mathbb{R}^d$-valued random vectors as the density of the measure ν is symmetric. Moreover we have

$$\|f\|_\infty \le \sum_{j=1}^{d}\Big(\Big\|\frac{\partial\Phi}{\partial\theta_j}(\theta_0,\cdot)\Big\|_{\nu;q} + \|\nu\|_\infty\Big)\cdot\big\|\Phi(\theta_0,\cdot)\big\|_{\nu;q} < \infty,$$

by the Cauchy-Schwarz inequality in the notation of Lemma 15. Furthermore note that

$$\int_{-1}^{1} E\big\{f(Y_1,X_1,Z_1)\mid Z_1 = z\big\}dz = 0\,,$$

thanks to (26) and (29). Then Lemma 16 provides that $\big(\sqrt{n}\Delta_n\big)_n$ converges in distribution to some centered Gaussian random variable with the covariance matrix $2\int_{-1}^{1}\sigma^2(z)/f_Z(z)dz$ where

$$\begin{aligned}
\sigma^2(z) = q^{-2}\sum_{k,k'=1}^{q}\iint &\nabla_\theta\Phi(\theta_0,t,I_k)\{\nabla_\theta\overline{\Phi(\theta_0,t',I_{k'})}\}' \\
&\cdot\operatorname{cov}\Big(\big\{\exp(-itY_1)1_{I_k}(X_1) - \phi(X_1)\overline{\Phi(\theta_0,t,I_k)}\big\}\big\{\exp(it'Y_1)1_{I_{k'}}(X_1) - \phi(X_1)\Phi(\theta_0,t',I_{k'})\big\}\Big|Z_1 = z\Big)\,dt\,dt'\,.
\end{aligned} \tag{78}$$

This also yields that $\Delta_n = \mathscr{O}_P(n^{-1/2})$.

As $\hat{\theta}_n$ minimizes the function (31) by definition we have that

$$0 = \Big\langle\frac{\partial}{\partial\theta_j}\Phi(\hat{\theta}_n,\cdot), \hat{\Phi}_n(\cdot) - \hat{a}_{I,n}\cdot\Phi(\hat{\theta}_n,\cdot)\Big\rangle_{\nu,q}, \qquad \forall j = 1,\dots,d\,. \tag{79}$$

Therein note that, thanks to Proposition 17 and the consistency part (33) of Theorem 9 as proved above, we may restrict to the events $\{\hat{a}_{I,n} > 0\}$ and $\{\hat{\theta}_n \in \dot{\Theta}\}$ where $\dot{\Theta}$ denotes the interior of the set Θ. Also those findings yield

that, by Taylor approximation,

$$\begin{aligned}\nabla_\theta \Phi\big(\hat\theta_n,\cdot\big) &= \nabla_\theta \Phi\big(\theta_0,\cdot\big) + \Big\{\frac{\partial^2 \Phi}{\partial\theta_j \partial\theta_{j'}}(\theta_0,\cdot)\Big\}_{j,j'=1,\ldots,d}(\hat\theta_n-\theta_0) + o_P(n^{-1/2})\,,\\ \Phi\big(\hat\theta_n,\cdot\big) &= \Phi\big(\theta_0,\cdot\big) + \big\{\nabla_\theta \Phi\big(\theta_0,\cdot\big)\big\}'(\hat\theta_n-\theta_0) + o_P(n^{-1/2})\,,\end{aligned}$$

so that (79) provides that

$$\hat a_{I,n} G(\hat\theta_n-\theta_0) = \Delta_n + o_P(n^{-1/2})\,.$$

Assumption 9 ensures the invertibility of the Gram matrix G, which completes the proof of the theorem. ∎