Supplemental Appendix to "Random Coefficients on Endogenous Variables in Simultaneous Equations Models"

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

July 13, 2017

Abstract

This supplemental appendix provides several results which complement those in the main paper. First I consider identification under unbounded support, rather than full support. Then I provide some follow-up derivations to several points discussed in the main text. Finally, I provide several additional estimation results. I first study several Monte Carlo simulations to examine the finite sample performance of the proposed estimator. I then discuss bandwidth selection and how to incorporate covariates in estimation. I conclude with a proof of consistency for the proposed nonparametric density estimator.

A Identification via unbounded support

In this section I show that the full support assumption used in theorem 3 can be relaxed to the following unbounded support assumption.[1]

Assumption A4$''$ (Instruments have unbounded support). $\text{supp}(Z_1 \mid X = x, Z_2 = z_2)$ is unbounded, for at least one element $z_2 \in \text{supp}(Z_2 \mid X = x)$, for each $x \in \text{supp}(X)$. Likewise, $\text{supp}(Z_2 \mid X = x, Z_1 = z_1)$ is unbounded, for at least one element $z_1 \in \text{supp}(Z_1 \mid X = x)$, for each $x \in \text{supp}(X)$.

Theorem S1. Under A1, A2, A3, and A4$''$, the conditional distributions $\gamma_1 \mid X = x$ and $\gamma_2 \mid X = x$ are identified for each $x \in \text{supp}(X)$.

Proof of theorem S1. By A1 we can solve for the reduced form equations

$$Y_1 = \frac{U_1 + \gamma_1 U_2 + \beta_1 Z_1 + \delta_1' X + \gamma_1 \delta_2' X}{1 - \gamma_1 \gamma_2} + \frac{\gamma_1 \beta_2 Z_2}{1 - \gamma_1 \gamma_2}$$

[1] I thank a referee for pointing out this result.

and

$$Y_2 = \frac{\gamma_2 U_1 + U_2 + \gamma_2 \beta_1 Z_1 + \gamma_2 \delta_1' X + \delta_2' X}{1 - \gamma_1 \gamma_2} + \frac{\beta_2 Z_2}{1 - \gamma_1 \gamma_2}.$$

Taking the ratio of these two equations yields

$$\begin{aligned} \frac{Y_1}{Y_2} &= \frac{(U_1 + \gamma_1 U_2 + \beta_1 Z_1 + \delta_1' X + \gamma_1 \delta_2' X) + \gamma_1 \beta_2 Z_2}{(\gamma_2 U_1 + U_2 + \gamma_2 \beta_1 Z_1 + \gamma_2 \delta_1' X + \delta_2' X) + \beta_2 Z_2} \\ &= \frac{(U_1 + \gamma_1 U_2 + \beta_1 Z_1 + \delta_1' X + \gamma_1 \delta_2' X)/(\beta_2 Z_2) + \gamma_1}{(\gamma_2 U_1 + U_2 + \gamma_2 \beta_1 Z_1 + \gamma_2 \delta_1' X + \delta_2' X)/(\beta_2 Z_2) + 1} \\ &\equiv \frac{\frac{V_1}{Z_2} + \gamma_1}{\frac{V_2}{Z_2} + 1}. \end{aligned}$$

A2 ensures that the ratio random variable in the second line is well defined. Note that $(V_1, V_2) \perp\!\!\!\perp Z_2 \mid Z_1, X$ by A3. Fix a $(z_1, x) \in \operatorname{supp}(Z_1, X)$. By A4$''$ there exists a sequence $\{z_{2s}\}_{s=1}^{\infty}$ with $z_{2s} \in \operatorname{supp}(Z_2 \mid Z_1 = z_1, X = x)$ such that $z_{2s} \to \pm\infty$ as $s \to \infty$. Hence for any $t \in \mathbb{R}$,

$$\begin{aligned} \lim_{s\to\infty} \mathbb{P}\left(\frac{Y_1}{Y_2} \leq t \mid X = x, Z_1 = z_1, Z_2 = z_{2s}\right) &= \lim_{s\to\infty} \mathbb{P}\left(\frac{V_1/z_{2s} + \gamma_1}{V_2/z_{2s} + 1} \leq t \mid X = x, Z_1 = z_1, Z_2 = z_{2s}\right) \\ &= \lim_{s\to\infty} \mathbb{P}\left(\frac{V_1/z_{2s} + \gamma_1}{V_2/z_{2s} + 1} \leq t \mid X = x, Z_1 = z_1\right) \\ &= \mathbb{P}(\gamma_1 \leq t \mid X = x, Z_1 = z_1) \\ &= \mathbb{P}(\gamma_1 \leq t \mid X = x). \end{aligned}$$

The left hand side is known from the data and hence the right hand side is point identified.

An analogous proof can be used to obtain the distribution of $\gamma_2 \mid X$. □

The proof of theorem 3 relied on the classical Cramér-Wold theorem. Theorem S1 provides an alternative and complementary identification proof which instead explicitly uses identification at infinity. This alternative proof allowed us to relax full support to unbounded support. In particular, this allows us to achieve point identification of the distributions of $\gamma_1 \mid X$ and $\gamma_2 \mid X_2$ when (1) the instrument is discretely distributed with unbounded support and (2) the instrument has unbounded support only on the positive part of the real line (or, similarly, only on the negative part). This identification at infinity proof strategy is illustrated more simply in the following result for single equation models.

Lemma S1. Suppose

$$Y = A + B'Z$$

where Y and A are scalar random variables and B and Z are random K-dimensional vectors. Suppose the joint distribution of (Y, Z) is observed. If $Z \perp\!\!\!\perp (A, B)$ and $\operatorname{supp}(Z_k \mid Z_{-k} = z_{-k})$ is unbounded for at least one element $z_{-k} \in \operatorname{supp}(Z_{-k})$, then the distribution of B_k is point identified.

Proof of lemma S1. We have

$$\frac{Y}{Z_k} = \frac{A}{Z_k} + B_k + B_{-k}' \frac{Z_{-k}}{Z_k}.$$

By assumption there is a sequence $\{z_{ks}\}_{s=1}^{\infty}$ with $z_{ks} \in \text{supp}(Z_k \mid Z_{-k} = z_{-k})$ such that $z_{ks} \to \pm\infty$ as $s \to \infty$. Hence for any $t \in \mathbb{R}$,

$$\begin{aligned} \lim_{s\to\infty} \mathbb{P}\left(\frac{Y}{Z_k} \leq t \mid Z_{-k} = z_{-k}, Z_k = z_{ks} \right) &= \lim_{s\to\infty} \mathbb{P}\left(\frac{A}{z_{ks}} + B_k + B_{-k}' \frac{z_{-k}}{z_{ks}} \leq t \mid Z_{-k} = z_{-k}, Z_k = z_{ks} \right) \\ &= \lim_{s\to\infty} \mathbb{P}\left(\frac{A}{z_{ks}} + B_k + B_{-k}' \frac{z_{-k}}{z_{ks}} \leq t \right) \\ &= \mathbb{P}(B_k \leq t). \end{aligned}$$

The left hand side is known from the data and hence the right hand side is point identified. □

Lemma S1 is similar to Beran and Hall (1992) theorem 2.1, except that they assumed $A \perp\!\!\!\perp B$. It is also similar to Beran and Millar (1994) proposition 2.2′, except that they assumed B had compact support and that (A, B) belonged to a known tight class of cdfs. Both of these previous results, however, obtained point identification of the joint distribution of (A, B) rather than just the marginal distribution of B_k.

B Auxiliary derivations

B.1 Derivations to show 2SLS estimates a weighted average effect parameter

Proof. We have

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

and

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

□

B.2 Equilibrium stability

In this section I consider the relationship between the parameters of the linear simultaneous equations model and stability of its equilibrium. Consider a single realization of the unobservables (Γ, B, D, U). Although there are many ways to model disequilibrium dynamics, consider the simple dynamic process

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

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

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

denote the equilibrium (or steady state) value of outcomes. Say this equilibrium is *globally stable* if for any $Y_0 \in \mathbb{R}^2$, $Y_t \to Y$ as $t \to \infty$. As mentioned in the paper, the equilibrium is globally stable if and only if $|\gamma_1 \gamma_2| < 1$. For reference, the following is a straightforward proof of this result.

Proof of global stability characterization. Let

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

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

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

which implies

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

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

B.3 Limited dependent variables

As noted on page 12 of the paper, bounded support of (Y_1, Y_2) ensures that A5 holds, but such a bound may introduce dependence between the instruments Z and the random coefficients to ensure that the bound is satisfied, depending on the support of Z. For bounded outcomes a limited dependent variable (LDV) type model may be more appropriate. In this case system (1) could be viewed as a system of latent endogenous variables with random coefficients on the endogenous variables, with our observable outcomes being a censored version of these latent variables. Identification of such random coefficient LDV models remains an open question. For important relevant work on simultaneous equation LDV models, see Matzkin (2012) and Blundell and Matzkin (2014).

In the empirical illustration of section 5, I follow Sacerdote (2001), who also did not use a limited dependent variable model in his constant coefficient analysis. While stopping short of a full LDV analysis, we can still perform some rough calculations to check whether the linear model is a reasonable approximation, despite the bounded support of the outcome variables. One way to do this is to use our estimates to generate predicted outcomes and check whether they violate the logical bounds of the data. Obtaining predicted outcome values is slightly complicated here since the values of the coefficients γ_i for each i are not identified; only their distribution is. However, to get a rough idea of whether predicted values violate the logical support constraints one can do the following: For each person, randomly draw a value γ_i from the estimated density $\widehat{f}_\gamma$. Then define

$$\widehat{\text{GPA}}_{it} = \widehat{\mu}_{\text{3SLS}} + \gamma_i \cdot \text{GPA}_{jt} + \widehat{\beta}_{\text{3SLS}} \cdot \text{GPA}_{i,t-1}.$$

Here j refers to person i's friend. Then compute the proportion of observations which fall outside the support of GPA, $[0,4]$. I repeated this 100 times. The average proportion outside the support is 36%. Many of those observations are close to the support—only 14% fall outside of $[0,5]$.

There are a few caveats to interpreting these numbers, however:

- They assume $U_{it} = 0$ for all i.
- They assume γ_i is independent of GPA_{jt}, which is not correct (because of simultaneity).
- They impose a constant coefficient on the instrument, $\beta_i = \widehat{\beta}_{\text{3SLS}}$.

With these caveats, and since most of the observations do not violate the constraint too badly, I interpret this as evidence that a linear approximation is roughly valid for illustrating the techniques of the paper. But these back-of-the-envelope calculations do suggest that it will be important to develop fully rigorous methods for extending the methods of this paper to handle limited dependent variables in future work.

B.4 On the potentially relevant friend assumption in theorem 6

In this section I first give a simple sufficient condition for assumption 6 in theorem 6. I then show that, in several special cases, assumption 6 is is equivalent to the assumption that each person has at least one friend: $N_i \geq 1$ for all $i = 1, \ldots, N$. I conjecture that this result is true in general, but I explain why obtaining a general proof is difficult.

Proposition S1. Suppose either of the following hold.

1. The joint distribution of reduced form random coefficients $(\pi_1, \ldots, \pi_N)$ has a density function.
2. The joint distribution of $(\gamma_1, \ldots, \gamma_N)$ has a density function.

Then assumption 6 holds.

The second condition in proposition 1 is weaker than the first since it does not constrain the distribution of the β_i or U_i terms.

Proof of proposition S1.

1. We want to show the term
$$1_{ij} + \sum_{k \neq i, k \neq j} 1_{ik}(\pi_{kj}/\pi_{jj}) \tag{S1}$$
is nonzero with probability one. Setting this equal to zero and multiplying by π_{jj} we obtain
$$1_{ij}\pi_{jj} + \sum_{k \neq i, k \neq j} 1_{ik}\pi_{kj} = 0.$$
The left hand side is a polynomial in the reduced form coefficients. Hence this equation defines a hypersurface and so has Lebesgue measure zero by lemma 3 in the main appendix 6. Since the reduced form coefficients have a density, they place zero mass on this hypersurface.

2. Recall that $\pi_{kj} = [A_n^{-1}]_{kj}\beta_j$, where A_n is defined as in the proof of theorem 6. A_n only depends on the γ_k terms. Since the β_j's cancel in the ratios of the reduced form coefficients, the distribution of equation (S1) only depends on the distribution of the γ_k terms. If we set equation (S1) equal to zero and rearrange slightly, we obtain
$$1_{ij}[A_n^{-1}]_{jj} + \sum_{k \neq i, k \neq j} 1_{il}[A_n^{-1}]_{kj} = 0.$$
The left hand side is a polynomial the γ_k terms, by equations (11′) and (12′). The proof now follows as in part 1.

□

Next we have the following simple proposition.

Proposition S2. Suppose assumption 6 holds. Then each person has at least one friend: $N_i \geq 1$.

Proof of proposition S2. This follows by contradiction: If $1_{ij} = 0$ for all j then equation (S1) equals zero regardless of the value of π_{kj}/π_{jj}, and hence assumption 6 would fail. □

This implication of assumption 6 is reasonable since if one person i was not influenced by anyone then $\gamma_i \mid G$ would be trivially unidentified because γ_i would not enter the outcome equation (8).

The following proposition shows that the converse holds in three special cases.

Proposition S3. Suppose assumptions 1–5 of theorem 6 hold. Suppose $N_i \geq 1$ for all $i = 1, \ldots, N$. Suppose further that one of the following special cases holds.

1. The network is complete: For all i, $1_{ij} = 1$ for all $j \neq i$.

2. There are three people in the network: $N = 3$.

3. N and the network G are arbitrary but we only allow nonnegative social interaction effects: There is a $\tau \in (0,1)$ such that $\mathbb{P}(0 \leq \gamma_i \leq \tau \mid G) = 1$ for all $i = 1, \ldots, N$.

Then assumption 6 holds.

Proof of proposition S3.

1. This is simply the case of theorem 5.

2. Recall the discussion following theorem 6. There we had

$$\gamma_2 \big[1_{21} + 1_{23}(\pi_{31}/\pi_{11}) \big] = N_2(\pi_{21}/\pi_{11})$$

   and

$$\frac{\pi_{21}}{\pi_{11}} = \frac{\gamma_2}{N_2} \cdot \frac{\left(1_{21} + 1_{23}1_{31}\frac{\gamma_3}{N_3} \right)}{1_{21} - 1_{23}1_{32}\frac{\gamma_2}{N_2}\frac{\gamma_3}{N_3}}.$$

   Hence equation (S1) can be written as

$$1_{21} + 1_{23}(\pi_{31}/\pi_{11}) = \frac{\left(1_{21} + 1_{23}1_{31}\frac{\gamma_3}{N_3} \right)}{1_{21} - 1_{23}1_{32}\frac{\gamma_2}{N_2}\frac{\gamma_3}{N_3}}.$$

   Let person 1 affect person 2, $1_{21} = 1$ (by assumption, each person has at least one friend). Consider this equation as a function of (γ_2, γ_3) and minimize its magnitude. Recall that $|\gamma_k| \leq \tau < 1$. The minimum magnitude is obtained at $\gamma_3 = -\tau$ and $\gamma_2 = \tau$. At this value, the denominator is some positive number and the numerator bounded below by $1 - \tau$, regardless of the rest of the network structure (besides $1_{21} = 1$, that is). Hence equation (S1) is bounded away from zero for all values of $(\gamma_2, \gamma_3) \in [-\tau, \tau]^2$. The same argument applies symmetrically to the other two people in the network.

3. Let A_n be defined as in the proof of theorem 6. By part (N38) of theorem 2.3 on pages 134–138 of Berman and Plemmons (1979), all elements of A_n^{-1} are nonnegative. Recall from the proof of theorem 6 that the diagonal elements of A_n^{-1} have the same structure as $\det(A_n)$, and A_n is always invertible, and therefore $\det(A_n) > 0$ always holds. This same argument can be used to show that the diagonal elements of A_n^{-1} are always strictly positive. Finally, since each person k has at least one person i who affects them, we have $1_{ki} = 1 > 0$ for that person. Thus, for this specific i,

$$1_{ki} + \sum_{j \neq k, j \neq i} 1_{kj}[A_n^{-1}]_{ji}/[A_n^{-1}]_{ii} > 0$$

   always holds, since this is 1 plus the sum of elements which are nonnegative. Multiplying by β_i/β_i inside the summation completes the proof.

□

I conjecture that the conclusion of proposition 3 holds in general, not just in these three special cases. Obtaining a theoretical proof for the general case, however, appears to be difficult. Since the result is true in the nonnegative γ_k case (case 3 above), the key difficulty is allowing for negative social interaction effects. For general network structures, this is an open problem in the microeconomic theory literature on network comparative statics. For example, in their recent survey of the literature Bramoullé and Kranton (2016) note that "there are few results that relate the aggregate impacts under substitutes to the structure of the network" and that the existing results focus on bipartite graphs. By 'substitutes' they mean negative social interaction effects and by 'aggregate impacts' they mean the equilibrium impact of changing one person's characteristic (their instrument) on another person's outcome; this is precisely what the reduced form coefficients capture. Advances in microeconomic theory may eventually allow us to confirm my conjecture above for the remaining open cases.

Even without a general proof, a direct proof based on the structure of equations (11′) and (12′), as in the 3 person network proof above, is in principle possible for networks of 4 or more people. Rather than doing this by hand, one could check this by numerically minimizing the polynomials over all γ_k's in $[-\tau, \tau]$ as in the 3 person argument above. This could be simplified by noticing that the optima occur at the vertices of the joint support of the γ_k's, as in the 3 person case. The difficulty with this approach is that this numerical optimization would have to be done for each possible network realization. And the number of possible networks, $2^{N(N-1)}$, grows extremely fast in the number of nodes N, making this impractical. In empirical applications, however, one would only have to check for the network structures actually observed in one's dataset, which could make this direct approach feasible.

I conclude this section by illustrating a subtlety related to the difficulty of obtaining a general proof of my conjecture above. Assumption 6 only requires that there exists at least one person j who is potentially relevant. For certain realizations of the structural random coefficients, it is possible that a person is affected by many people, and yet some of them are not potentially relevant.

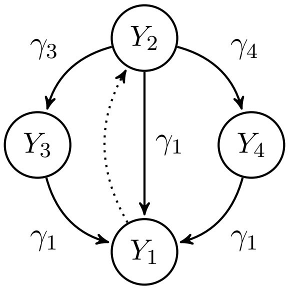


Figure 1: An example network structure which can lead to cancellation of network effects.

For example, consider the 4 person network in figure 1. Here the reduced form effect of 2's instrument on 1 is

$$\frac{\beta_2}{\det(\widetilde{\Gamma})}[\gamma_1(1 + \gamma_3 + \gamma_4)].$$

This effect depends on the sum of the three paths from 2 to 1: the direct path effect, γ_1, and the two indirect paths through persons 3 and 4, $\gamma_1\gamma_3$ and $\gamma_1\gamma_4$. A problem occurs if, for example, $\gamma_3 = \gamma_4 = -0.5$. This is allowed under any reasonably large τ, like $\tau = 0.95$. In this case, the reduced form effect of 2's instrument on 1 zero, regardless of the value of γ_1. Thus 2 is not a potentially relevant friend for 1: The positive direct effect of 2 on 1 was exactly canceled out by the negative indirect effects through 3 and 4. For such a network structure, the solution to identify γ_1 is simply to look at a different friend's effect on 1. For example, the reduced form effect of 3's instrument on 1 is just

$$\frac{\beta_3}{\det(\widetilde{\Gamma})} \cdot \gamma_1$$

since there is only one path from 3 to 1. Hence $\pi_{31}/\pi_{11} = \gamma_1$.

This example was merely to illustrate the difficulties that arise with negative social interaction effects in networks. As shown at the beginning of this section, if the γ_k's have a joint density, then this exact cancellation happens with probability zero for all friends and hence all of one's direct friends are potentially relevant. Nevertheless, even if one does not wish to make this distributional assumption on the γ_k's, my conjecture above is that, regardless of the distribution of the γ_k's (subject to assumptions 1–5), it is never possible for *all* of one's friends to be irrelevant at the same time; one of them must be.

C Additional estimation results

In this section I present several Monte Carlo simulations, discuss bandwidth selection, discuss a dimension reduction method when including covariates in estimation, and give a consistency result.

C.1 Monte Carlo simulations

To examine the nonparametric estimator's finite sample performance, I run several Monte Carlo simulations. The conditions of both theorems 2 and 3 hold in all simulations so that either result could be used to ensure identification. I consider seven different generating processes.

In the first set of dgps, I hold everything constant except two aspects. First, the common marginal distribution f_γ is one of the following:

1. A truncated normal with pre-truncation mean 0.4 and standard deviation 0.05.
2. A Beta distribution with shape parameter 6 and scale parameter 3.
3. A truncated mixture of two normals: 0.3 weight on $\mathcal{N}(0.3, 0.05)$ and 0.7 weight on $\mathcal{N}(0.7, 0.06)$.

See figure 2 for plots of these marginal distributions. The original support of these three densities is $[0, 1]$, which is then scaled to $[0, 0.95]$, which helps ensure that f_γ is identified. Second, the instruments Z_1 and Z_2 are either standard Cauchy or $\mathcal{N}(0, 3)$ distributed. By varying over the three choices of f_γ and the two choices for the distribution of the instruments we obtain the first six dgps.

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

The parameters for the seventh dgp were chosen to reflect the empirical illustration in section 5 of the paper. Specifically, the instruments Z_1 and Z_2 are $\mathcal{N}(2.8, 0.8)$, which approximates the amount of variation in the instruments in the data. The own coefficient on the instrument is 1, which reflects a plausible value of this coefficient in the application, especially given the reduced form coefficient in table 2 of the paper. (U_1, U_2) are bivariate normal with $\mu_u = 0$, $\sigma_u = 0.8$, and $\rho_u = 0$, and the intercept is -2. γ_1 and γ_2 are independent and f_γ is chosen to be the bimodal distribution number 3 above, which approximates the positive portion of the estimated distribution plotted in figure 1 in the paper (in these simulations, I restrict attention to nonnegative γ for simplicity). These choices all together imply that, based on estimating each reduced form mean regression separately without imposing cross equation constraints, the first stage F-statistic for assessing instrument strength is on average 251, which is between the values of 372 and 192 obtained in the empirical illustration. These choices also imply that, although the outcome variables are jointly bivariate normal, most of their mass is close to $[0, 4]$, the support of the outcome variables in the application.

For each dgp, I compute several statistics. First, I compute the median bias of several scalar parameter estimators. For any scalar parameter κ, the estimated median bias is the median of $\hat{\kappa}_s - \kappa$ over all $s = 1, \ldots, S$, where S is the total number of Monte Carlo simulations, and s indexes each simulation run. As a measure of spread, I compute the estimated interquartile range (IQR) as the interquartile range of $\hat{\kappa}_s - \kappa$ over all simulations s. I use the median bias and IQR here because the distributions of the 2SLS and RC estimators may have thick tails. I use $S = 1000$ simulations, which yields simulation standard errors small enough to make statistically significant comparisons. I compute these statistics for the nonparametric estimator of the random coefficients' mean:

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

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

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

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

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

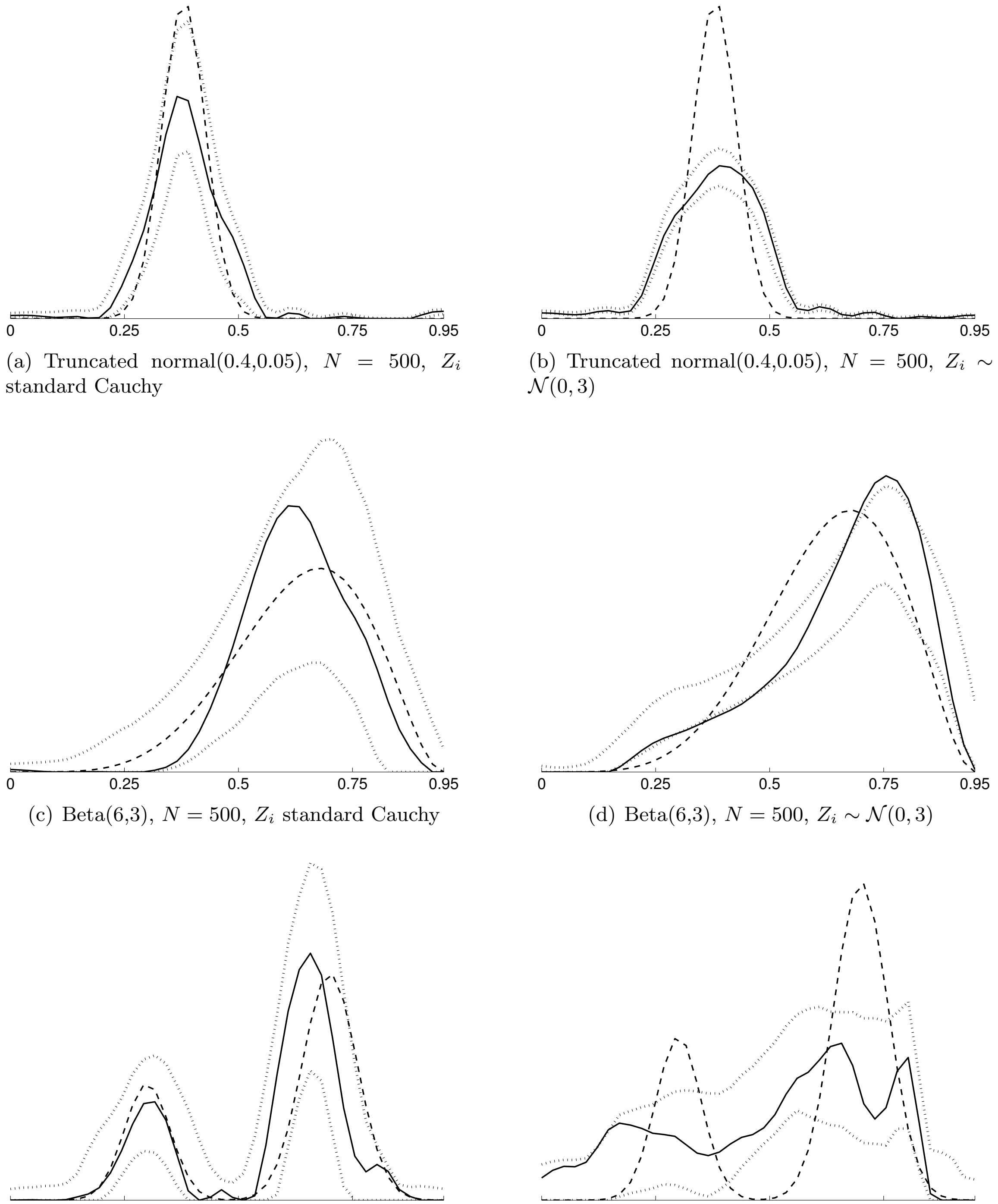
The median ISE is estimated by the median of this value over all simulations.

Figure 2 shows example plots of $\widehat{f}_\gamma$ versus the true density, for $N = 500$ and the first six dgps. The estimator recovers the general shape of the true density in five of the six dgps, although it performs better with Cauchy distributed instruments compared to the normally distributed instruments. This is to be expected given the previous literature on nonparametric estimation in single equation random coefficients models. As discussed in section 4 of the paper, the only two options for the first step of my estimator are Beran, Feuerverger, and Hall (1996) and Hoderlein, Klemelä, and Mammen (2010). The assumptions in Beran et al. (1996) require thicker than normal tailed regressors. They also show that the rate of convergence depends on the rate at which the density of the regressors goes to zero in the tails: the thinner the regressor tails, the slower the rate. Likewise, the main theory of Hoderlein et al. (2010) also requires thicker than normal tailed regressors (see their theorem 3, however, where they show one way to relax this assumption). This property affects the first step of my estimator, and hence carries through to the final step estimator of $\widehat{f}_\gamma$, as we can see in the plots.

For the bimodal f_γ with normally distributed instruments, $\widehat{f}_\gamma$ does a poor job estimating the shape of the true density, although it does generally recover the feature that there is more mass above 0.5 than below. There are two reasons for this poor performance. The first is the same issue with normal tailed regressors mentioned above. The simulations suggest, however, that this difference in tail thickness matters much more for bimodal than unimodal distributions of f_γ. The second reason is that the linear combination reduced form pdf being estimated in step one often has many modes with widely differing levels of local spread. For example, figure 4 shows two such pdfs. In contrast, the two unimodal distributions of f_γ guarantee that the linear combination pdfs always have a single mode, which makes them easier to estimate.[2] Figure 4 shows two such pdfs for a unimodal f_γ. Figure 3 shows an example plot of $\widehat{f}_\gamma$ versus the true density for $N = 500$ and the empirically calibrated seventh dgp. Similar comments apply, although in this case the shape is somewhat better estimated. Although the shape is difficult to estimate for bimodal f_γ, we will see next that some functionals of the density are well estimated.

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

[2] Allowing for local bandwidths in this first step (for example, Scott (2015) section 6.6) may improve the performance of $\widehat{f}_\gamma$ in the bimodal case. Such local bandwidth selection for single equation random coefficient estimation has not been previously studied, however, so I leave this point to future work.



(a) Truncated normal(0.4,0.05), $N = 500$, Z_i standard Cauchy

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

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

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

(e) Bimodal, $N = 500$, Z_i standard Cauchy

(f) Bimodal, $N = 500$, $Z_i \sim \mathcal{N}(0, 3)$

Figure 2: Nonparametric estimates of f_γ, the common marginal distribution of random coefficients. Thick dotted lines show the true density, solid lines show the estimated density, and thin dotted lines show 95% pointwise Monte Carlo confidence bands. Estimates correspond to the simulation with integrated squared error at the median over all simulations.

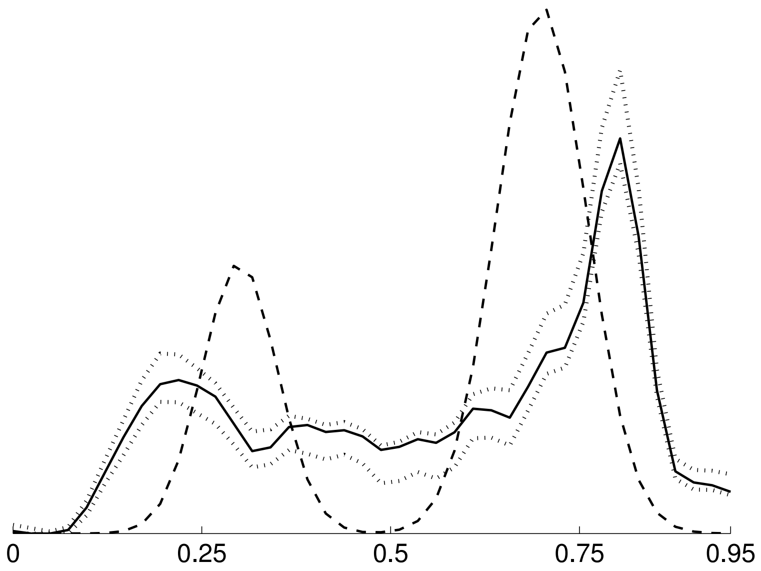


Figure 3: Nonparametric estimate of f_γ for the empirical bimodal dgp, $N = 500$, $Z_i \sim \mathcal{N}(2.8, 0.8)$. Estimate corresponds to the simulation with integrated squared error at the median over all simulations. Thick dotted lines show the true density, solid lines show the estimated density, and thin dotted lines show 95% pointwise Monte Carlo confidence bands.

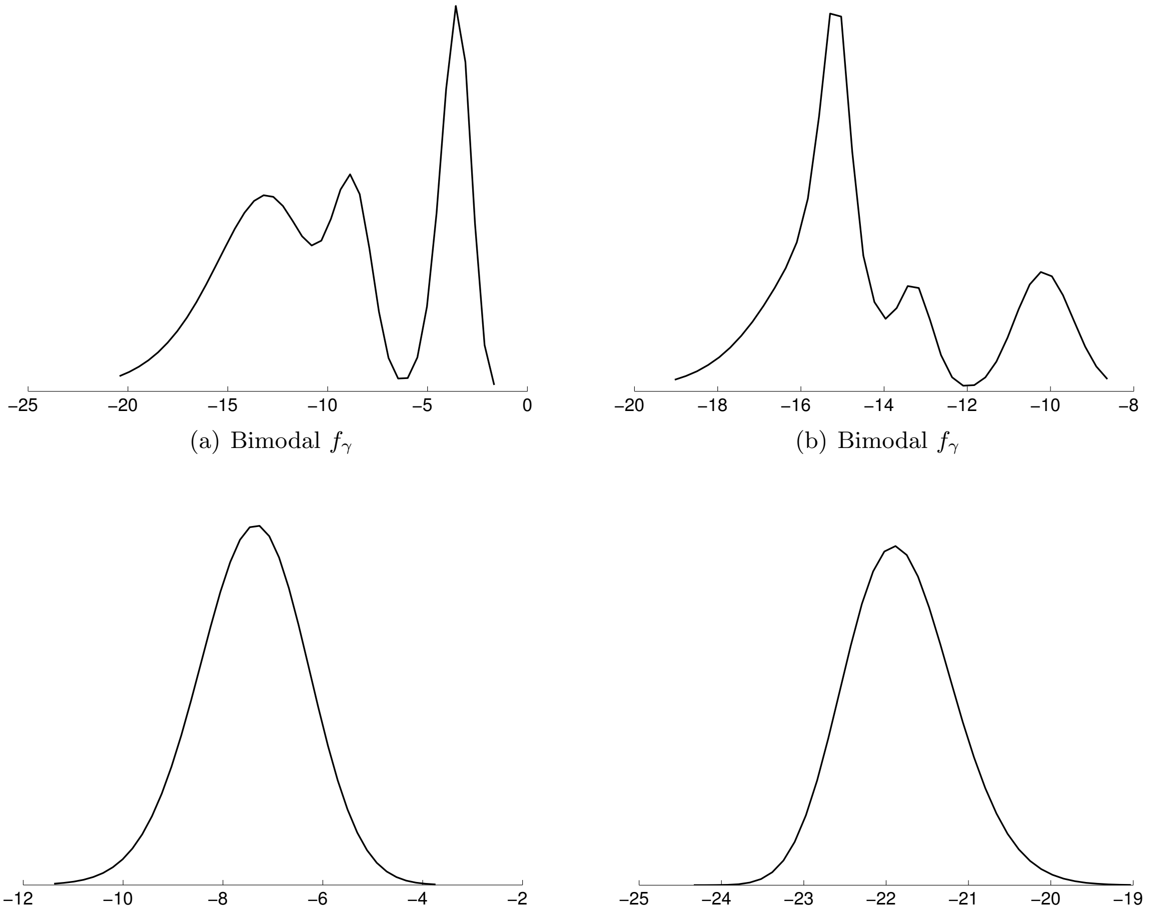
of the ISE. Table 1 shows results for Cauchy distributed instruments while table 2 shows results for normally distributed instruments.

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

The second dgp is slightly asymmetric and more spread out than the first dgp. 2SLS and the RC mean estimator continue to estimate $\mathbb{E}(\gamma)$ well, although not as well as in the first dgp. Again the 2SLS distribution has a substantially larger spread. The RC median estimator has a smaller bias than the RC mean estimator. The RC IQR estimator performs well in this dgp, with a median bias and spread one order of magnitude smaller than the truth.

The third dgp is the bimodal distribution. As with the other dgps, the RC mean and median estimators perform well. 2SLS, however, performs much worse than the RC estimator, with a larger bias and spread. The RC IQR estimator performs reasonably well, as in the second dgp.

Next consider table 2, with normally distributed instruments. Consider the first dgp. Despite the problems mentioned earlier with relatively thin tailed regressors, the RC estimators of the mean and median do very well. The RC estimators of the location of the distribution are comparable to 2SLS, which now also performs well with both a small bias and a small spread. The RC IQR



(a) Bimodal f_γ

(b) Bimodal f_γ

(c) Truncated normal(0.4,0.05) f_γ

(d) Truncated normal(0.4,0.05) f_γ

Figure 4: Several examples of the pdf of the linear combination random variable $t_1\pi_{12} + t_2\pi_{22}$.

Table 1: Monte Carlo results: Cauchy Z

	$\widehat{\mathbb{E}}(\gamma)$		$\widehat{\text{Med}}(\gamma)$	$\widehat{\text{IQR}}(\gamma)$	Median ISE
	2SLS	RC	RC	RC	RC
Truncated normal(0.4,0.05)	$\mathbb{E}(\gamma) = 0.38$		$\text{Med}(\gamma) = 0.38$	$\text{IQR} = 0.0641$	
$N = 500$	0.0025	0.0036	0.0013	0.0296	0.5574
	[0.0391]	[0.0079]	[0.0091]	[0.0175]	[0.3899]
$N = 1000$	0.0017	0.0033	0.0019	0.0244	0.4269
	[0.0390]	[0.0058]	[0.0067]	[0.0133]	[0.2716]
Beta(6,3)	$\mathbb{E}(\gamma) = 0.63$		$\text{Med}(\gamma) = 0.6455$	$\text{IQR} = 0.202$	
$N = 500$	0.0179	-0.0131	-0.0073	-0.0162	0.1732
	[0.1092]	[0.0352]	[0.0450]	[0.0547]	[0.1885]
$N = 1000$	0.0193	-0.0126	-0.0069	-0.0164	0.1398
	[0.1125]	[0.0257]	[0.0316]	[0.0463]	[0.1448]
Bimodal	$\mathbb{E}(\gamma) = 0.551$		$\text{Med}(\gamma) = 0.6327$	$\text{IQR} = 0.355$	
$N = 500$	0.0533	-0.0095	-0.0063	-0.0211	0.7435
	[0.1600]	[0.0308]	[0.0256]	[0.0445]	[0.6031]
$N = 1000$	0.0519	-0.0187	-0.0079	0.0027	0.7708
	[0.1552]	[0.0308]	[0.0207]	[0.0421]	[0.4327]

For each dgp: Median bias is first. IQR in brackets.

Table 2: Monte Carlo results: Normal Z

	$\widehat{\mathbb{E}}(\gamma)$		$\widehat{\text{Med}}(\gamma)$	$\widehat{\text{IQR}}(\gamma)$	Median ISE
	2SLS	RC	RC	RC	RC
Truncated normal(0.4,0.05)	$\mathbb{E}(\gamma) = 0.38$		$\text{Med}(\gamma) = 0.38$	$\text{IQR} = 0.0641$	
$N = 500$	0.0013	0.0052	0.0045	0.0716	1.5958
	[0.0070]	[0.0066]	[0.0074]	[0.0090]	[0.2446]
$N = 1000$	0.0008	0.0052	0.0041	0.0717	1.5865
	[0.0053]	[0.0053]	[0.0054]	[0.0076]	[0.2033]
Beta(6,3)	$\mathbb{E}(\gamma) = 0.63$		$\text{Med}(\gamma) = 0.6455$	$\text{IQR} = 0.202$	
$N = 500$	0.0230	0.0072	0.0318	0.0674	0.2300
	[0.0186]	[0.0209]	[0.0238]	[0.0358]	[0.1198]
$N = 1000$	0.0229	0.0060	0.0308	0.0674	0.2175
	[0.0132]	[0.0151]	[0.0177]	[0.0251]	[0.0783]
Bimodal	$\mathbb{E}(\gamma) = 0.551$		$\text{Med}(\gamma) = 0.6327$	$\text{IQR} = 0.355$	
$N = 500$	0.0259	-0.0055	-0.0545	-0.0636	1.2846
	[0.0213]	[0.0360]	[0.0384]	[0.0694]	[0.4133]
$N = 1000$	0.0276	-0.0036	-0.0545	-0.0696	1.3329
	[0.0140]	[0.0317]	[0.0332]	[0.0645]	[0.3934]
Empirical Bimodal	$\mathbb{E}(\gamma) = 0.551$		$\text{Med}(\gamma) = 0.6327$	$\text{IQR} = 0.355$	
$N = 500$	0.0267	0.0240	0.0230	0.0850	1.5186
	[0.0503]	[0.0172]	[0.0306]	[0.0355]	[0.1175]

For each dgp: Median bias is first. IQR in brackets.

estimator performs worse than in table 1. It is two times larger than the true IQR on average. This can also be seen in figure 2. In the second dgp the RC median and RC IQR estimates perform worse, while the RC mean estimates have smaller bias and comparable spread. As in the first dgp, 2SLS again has a comparable spread to the RC mean estimator.

Consider the third dgp. 2SLS performs similarly to the second dgp. As with the second dgp, the performance of the RC mean estimator is not much affected by the tail behavior of the instruments. The RC median and RC IQR estimates are worse than in table 1, however. Nonetheless, these RC median and RC IQR estimates still perform decently, with bias and spread an order of magnitude smaller than the truth. In particular, notice that the RC mean, median, and IQR estimators performance is similar across the second and third dgps. Finally, consider the fourth dgp, where f_γ continues to be bimodal as in the third dgp, but the remainder of the parameters have also been calibrated to the empirical illustration. Here the RC median and IQR estimates perform similarly to the third dgp. The RC mean estimator is worse, but is still comparable to 2SLS. Thus, for the third and fourth dgps, we see that although the overall shape is difficult to estimate well (as illustrated in figures 2 and 3), functionals of the pdf are still usefully estimated.

Overall, the simulation results suggest that the RC estimator often performs well with practical sample sizes. As expected, thicker instrument tails lead to better performance. The shape of true unimodal distributions is easier to estimate than the shape of multimodal distributions, while some functionals of these distributions are reasonably well estimated in both cases. This includes the spread functional IQR. In practice, empirical researchers may suspect multimodality to arise, for example, as a mixture of a discrete set of types, with each type having a unimodal distribution. If these types represent observed heterogeneity, such as men versus women, then the simulation findings suggest that accounting for this observed heterogeneity will improve estimator performance. For example, a simple approach would be to estimate separate pdfs f_γ for men and for women. I leave exploring more sophisticated approaches, as well as theoretical analysis of the estimators under unimodality versus multimodality assumptions, to future work.

The RC identification and estimation results studied in this paper and illustrated in this section provide a comprehensive view of the distribution of γ. Even when the entire shape is difficult to estimate, the location and spread are reasonably estimated. In contrast, traditional analysis based on the 2SLS estimand necessarily provides a limited summary of the distribution of γ.

C.2 Bandwidth selection

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

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

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

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

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

Notice that in all of the dgps with normally distributed instruments (plots (b), (d), and (f) of figure 2, and figure 3), the chosen tuning parameters appear to emphasize variance reduction at the cost of larger bias. With Cauchy distributed instruments, however, the bias and variance appear to be much better balanced (plots (a), (c), and (e) of figure 2). All dgps use the same procedure for choosing the bandwidth (discussed above), so this suggests that the quality of the bandwidth selector depends on the variation in the instrument. This could be explored further even in the single-equation exogenous regressor setting, where there is almost no work on bandwidth selection and where few simulation studies have been done. I leave this to future work.

I conclude this section by briefly discussing the chosen bandwidths in the simulations and the empirical application. Above I discussed the choice of bandwidth for the first step estimator. In principle, similar considerations can be applied to the tuning parameters δ and M. I leave a full analysis of how to select these tuning parameters to future work. For each simulation dgp, I pick a single M and δ, which are constant across simulation datasets. First, I let δ be smaller than machine zero. Second, I choose M so that when the true first stage density is used rather than the estimated density, we accurately recover the true f_γ. $M = 10$ for dgp 1, $M = 4$ for dgp 2, and $M = 8$ for dgp 3. $M = 30$ for the empirically calibrated dgp. I do not report the first step bandwidths for the simulations because there is a different bandwidth for each (t_1, t_2), as discussed above. In the empirical illustration of section 5 in the paper, I again let δ be smaller than machine zero. I choose M analogously to the scale a for the first step estimator bandwidths, by visual inspection. $M = 10$ and $a = 0.1$ for the empirical illustration. Again, these approaches certainly can be improved on, but I leave that to future work.

Table 3: Sensitivity of random coefficient estimates to choice of M

	$M = 8$	$M = 10$	$M = 12$
$\mathbb{E}(\gamma)$	0.3007	0.4411	0.4819
	[0.2634, 0.4217]	[0.3069, 0.5396]	[0.4489, 0.5656]
$Q_\gamma(0.25)$	0.0884	0.2482	0.2729
	[-0.0214, 0.2178]	[-0.0960, 0.3243]	[0.2520, 0.3414]
Med(γ)	0.3718	0.5981	0.6039
	[0.3580, 0.6628]	[0.3899, 0.6861]	[0.5416, 0.6856]
$Q_\gamma(0.75)$	0.7047	0.7104	0.7180
	[0.7009, 0.8045]	[0.6994, 0.8055]	[0.7009, 0.8473]
$\mathbb{P}(\gamma \geq 0)$	0.7759	0.8761	0.9054
	[0.7482, 0.8125]	[0.7282, 0.9020]	[0.8648, 0.9194]
Observations	330	330	330

Observations are pairs of best friends. 95% confidence intervals shown in brackets.

Tables 3 and 4 explore the sensitivity of the empirical estimates to bandwidth choice. Both tables report columns of the estimates in the last column of table 3 in the paper for different choices of M and a. Specifically, table 3 reproduces the main empirical results in the center column, obtained with $M = 10$ and $a = 0.1$, along with results for bandwidths M 20% larger and 20% smaller: $M = 8$ and $M = 12$. Table 4 is analogous, but for a choice of a 20% larger and 20% smaller: $a = 0.08$ and $a = 0.12$. Figure 5 shows the corresponding estimated densities for these different choices of bandwidths. Overall we see that varying the bandwidth up or down by 20% does not dramatically change the qualitative features of the estimates.

C.3 Incorporating covariates in estimation

While the above procedure can be extended immediately to allow for additional covariates X, this would involve estimating $1 + d_{Z_1} + d_{Z_2} + d_X$ dimensional joint density functions in the first step. In some cases it is possible to reduce this dimension to $1 + d_{Z_1} + d_{Z_2} + 1$ by constructing a single index for the additional covariates.

Table 4: Sensitivity of random coefficient estimates to choice of a

	$a = 0.08$	$a = 0.1$	$a = 0.12$
$\mathbb{E}(\gamma)$	0.4446 [0.0486, 0.5088]	0.4411 [0.3069, 0.5396]	0.4540 [0.4359, 0.4744]
$Q_\gamma(0.25)$	0.2330 [-0.5958, 0.3675]	0.2482 [-0.0960, 0.3243]	0.2577 [0.2444, 0.2710]
$\text{Med}(\gamma)$	0.6210 [0.0618, 0.6752]	0.5981 [0.3899, 0.6861]	0.6077 [0.5962, 0.6224]
$Q_\gamma(0.75)$	0.7731 [0.6267, 0.8040]	0.7104 [0.6994, 0.8055]	0.7123 [0.7066, 0.7341]
$\mathbb{P}(\gamma \geq 0)$	0.8243 [0.5119, 0.8900]	0.8761 [0.7282, 0.9020]	0.8934 [0.8742, 0.8975]
Observations	330	330	330

Observations are pairs of best friends. 95% confidence intervals shown in brackets.

For example, suppose δ_1 and δ_2 are degenerate on a common constant coefficient δ. This symmetry restriction is a reasonable assumption in social interactions applications, where the labels of person 1 versus person 2 do not matter. Suppose also that $d_{Z_1} = d_{Z_2} = 1$. Then the reduced form system is

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

which after defining some notation we write as

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

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

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

by taking the coefficient on X in a linear mean regression of Y_1 on $(1, X, Z_1, Z_2)$. $\tilde{\delta}$ is not quite

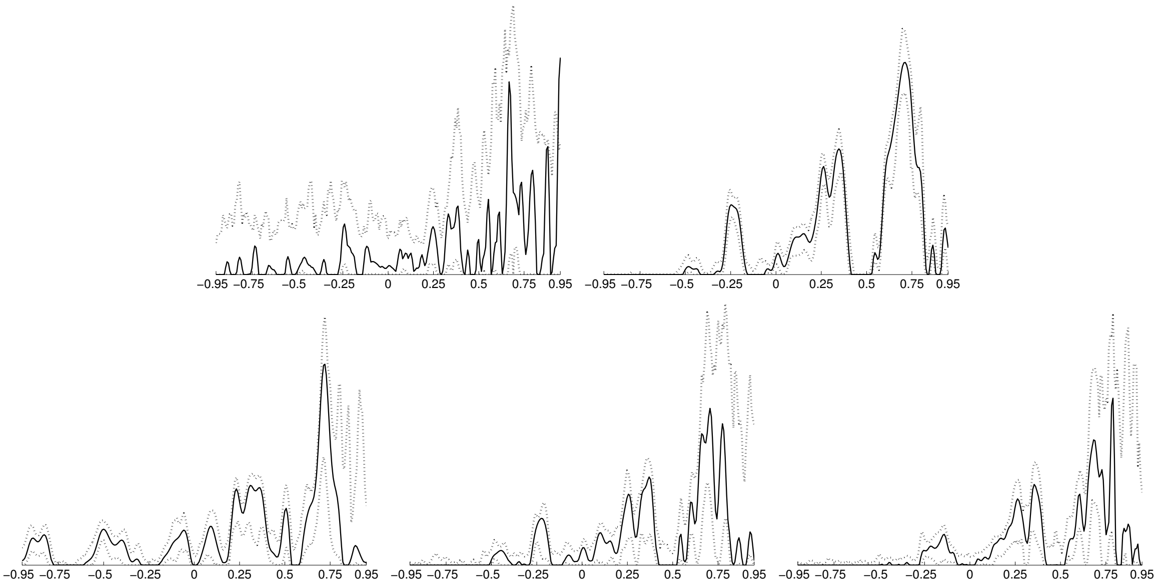


Figure 5: Sensitivity of density estimator to choice of M and a. Bottom center: $M = 10$, $a = 0.1$ (baseline estimator). Bottom left: $M = 8$, $a = 0.1$. Bottom right: $M = 12$, $a = 0.1$. Top left: $M = 10$, $a = 0.08$. Top right: $M = 10$, $a = 0.12$.

equal to δ because of the $\mathbb{E}[(1 + \gamma_1)/(1 - \gamma_1\gamma_2)]$ scale factor. Nonetheless, we now have the system

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

where the single index $\tilde{\delta}' X$ is estimated in the preliminary linear regression step. Thus, when estimating this system in step 1 by a single equation random coefficient estimator, we still obtain consistent estimates of the distribution of $t_1\pi_{12} + t_2\pi_{22}$ as needed. A similar dimension reduction can be done when a vector of covariates enters only one of the structural equations.

For estimating single equation random coefficient models with many covariates, Hoderlein et al. (2010) proposed assuming the coefficients on some of the covariates were constant, estimating them by a preliminary linear regression, and then partialing them out as in partially linear models. This approach does not work here because the determinant term $1/(1 - \gamma_1\gamma_2)$ ensures that all of the reduced form coefficients are random. Consequently, subtracting $\mathbb{E}(\tilde{\pi}_{1x}\delta')X$ from both sides of the reduced form equation does not remove the X term from the right hand side as it does in single equation models.

C.4 Consistency

In this section I give conditions under which the estimator of section 4 in the paper is consistent.

Assumption E1 (Conditions on the distribution of the reduced form random coefficients).

1. The reduced form random coefficients (π_1, π_2) have a density function.

2. The reduced form random coefficients (π_{12}, π_{22}) have compact support.

3. The density $f_{\pi_{12},\pi_{22}}$ is continuously differentiable.

4. The characteristic function of (π_{12}, π_{22}) is absolutely integrable: $\int_{\mathbb{R}^2} |\phi_{\pi_{12},\pi_{22}}(t_1, t_2)| \, dt_1 dt_2 < \infty$.

E1.4 is widely assumed in the literature when estimators depend on Fourier inversion, as in single equation random coefficient models or measurement error deconvolution. E1.4 implies that $f_{\pi_{12},\pi_{22}}$ is continuous on $\mathbb{R}^2$, which rules out densities which have jumps at the boundary, like the uniform distribution (Pinsky 2002 proposition 2.2.37, page 104). A simple sufficient condition for E1.4 is that $f_{\pi_{12},\pi_{22}}$ has third order partial derivatives and has a bounded Sobolev L_1-norm using third order derivatives (Pinsky 2002 exercise 2.2.39, page 105; Liflyand, Samko, and Trigub 2012 theorem 6.3 page 23 give a version using the Sobolev L_2-norm). Liflyand et al. (2012) give a comprehensive overview of necessary and sufficient conditions for E1.4.

All parts of E1 concern the distribution of reduced form coefficients. It is straightforward to give sufficient conditions for this assumption in terms of the structural coefficients, similarly to propositions 1 and 2 in the paper.

Let $\text{supp}[\Pi_2(t_1, t_2)] \subseteq [-B(t_1, t_2), B(t_1, t_2)]$, where $B(t_1, t_2) < \infty$ for all t_1, t_2. Such a $B(t_1, t_2)$ exists under assumption E1.2 that (π_{12}, π_{22}) has compact support.

Assumption E2 (First stage estimator uniform convergence). For any fixed $M > \delta > 0$,

$$\sup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2} \; \sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s; n) - f_{\Pi_2(t_1,t_2)}(s)| \to 0 \qquad a.s.$$

as $n \to \infty$.

E2 is a high-level assumption on the first stage single equation random coefficient estimator. Holding (t_1, t_2) fixed, rather than taking a supremum over it, E2 is just a standard sup-norm convergence condition for a density estimator.[3] The difference between this standard uniform convergence condition holding (t_1, t_2) fixed and the condition of E2 which also sups over (t_1, t_2) is relatively minor, since t_1 and t_2 typically enter the the kernel function in a qualitatively similar way as the argument s. E2 can be verified once one picks a specific first stage estimator. Below I sketch a verification of E2 for the inverse Radon transform estimator of Hoderlein et al. (2010).

Theorem S2. Suppose E1 and E2 hold. Then there exist sequences $M = M_n$ and $\delta = \delta_n$ with $M > \delta > 0$ and $M \to \infty$ and $\delta \to 0$ as $n \to \infty$ such that

$$\|\widehat{f}_{\gamma_2} - f_{\gamma_2}\|_\infty \to 0 \qquad a.s.$$

[3] Recall also that it is possible to achieve uniform convergence over the entire support, including the boundary, because E1.4 rules out jumps at the boundary.

as $n \to \infty$.

This result does not specify the required rates of the tuning parameters M and δ. I leave this question along with a further examination of the asymptotic properties of this estimator to future work.

Proof of theorem S2. This proof has three parts:

1. The first two parts concern the third step estimator of the joint pdf of (π_{12}, π_{22}),
$$\widehat{f}_{\pi_{12},\pi_{22}}(p_1, p_2; n, M, \delta).$$
Here I write this estimator as a function of the sample size and the tuning parameters for clarity. I first show that for any $\varepsilon > 0$, there exists an $M(\varepsilon)$, a $\delta(\varepsilon)$, and an $N(\varepsilon, M(\varepsilon), \delta(\varepsilon))$ such that for all $n \geq N(\varepsilon, M(\varepsilon), \delta(\varepsilon))$,
$$\sup_{(p_1,p_2)\in\mathbb{R}^2} |\widehat{f}_{\pi_{12},\pi_{22}}(p_1, p_2; n, M(\varepsilon), \delta(\varepsilon)) - f_{\pi_{12},\pi_{22}}(p_1, p_2)| \leq \varepsilon \qquad a.s.$$
That is, when we fix the tuning parameters at $M(\varepsilon)$ and $\delta(\varepsilon)$, the estimation error in the third step estimator is uniformly at most ε, almost surely, so long as n is sufficiently large.

2. I then show that part 1 implies that there exists a sequence of tuning parameters (M_n, δ_n) such that
$$\sup_{(p_1,p_2)\in\mathbb{R}^2} |\widehat{f}_{\pi_{12},\pi_{22}}(p_1, p_2; n, M_n, \delta_n) - f_{\pi_{12},\pi_{22}}(p_1, p_2)| \to 0 \qquad a.s.$$
as $n \to \infty$. That is, along this specific sequence of tuning parameters, the third step estimator is consistent in the sup-norm.

3. Finally, I show that sup-norm consistency of the third step estimator (using the same special sequence of M_n and δ_n from part 2) implies sup-norm consistency of the final step estimator.

Although the formal proof of part 2 below is long, the intuition is straightforward. From part 1 we see that any accuracy ε can be achieved by an estimator with a sufficiently large M and sufficiently small δ. A smaller ε will require a larger M and a smaller δ. Consider an arbitrary sequence $\varepsilon_{\widetilde{n}}$ going to zero. To each $\varepsilon_{\widetilde{n}}$ we associate an $M(\varepsilon_{\widetilde{n}})$ and a $\delta(\varepsilon_{\widetilde{n}})$. We thus obtain a triangular array like setup: For each $\widetilde{n}$, we consider the fixed accuracy $\varepsilon_{\widetilde{n}}$, and the fixed tuning parameters $M(\varepsilon_{\widetilde{n}})$ and $\delta(\varepsilon_{\widetilde{n}})$. Then, for these fixed parameters, we let the sample size n grow. Eventually, by part 1, we will achieve the desired accuracy. The only issue here is that the point at which we achieve the desired accuracy, $N(\varepsilon_{\widetilde{n}}, M(\varepsilon_{\widetilde{n}}), \delta(\varepsilon_{\widetilde{n}}))$ might be larger than $\widetilde{n}$. This is illustrated in the first part of table 5, which shows one possible outcome for an arbitrarily chosen sequence $\varepsilon_{\widetilde{n}}$. The solution is to just relabel the indices of our sequence $\varepsilon_{\widetilde{n}}$ appropriately, to obtain a table such as the second one in table 5. This relabeling will change our choice of tuning parameters $M(\varepsilon_{\widetilde{n}})$ and $\delta(\varepsilon_{\widetilde{n}})$. Hence this relabeling reflects the rate constraint on the tuning parameters: $M(\varepsilon_{\widetilde{n}})$

Table 5: Does $n \geq N(\varepsilon_{\widetilde{n}}, M(\varepsilon_{\widetilde{n}}), \delta(\varepsilon_{\widetilde{n}}))$? 1 if yes. 0 if no. Left: An arbitrary sequence $\varepsilon_{\widetilde{n}}$. Right: A specially chosen sequence $\varepsilon_{\widetilde{n}}$.

				n			
		1	2	3	4	5	$\cdots$
	1	0	0	1	1	1	
	2	0	0	1	1	1	
$\widetilde{n}$	3	0	0	0	1	1	
	4	0	0	0	0	1	
	5	0	0	0	0	0	
	6	0	0	0	0	0	
	$\vdots$						$\ddots$

				n			
		1	2	3	4	5	$\cdots$
	1	1	1	1	1	1	
	2	0	1	1	1	1	
$\widetilde{n}$	3	0	0	1	1	1	
	4	0	0	0	1	1	
	5	0	0	0	0	1	
	6	0	0	0	0	0	
	$\vdots$						$\ddots$

cannot grow too fast and $\delta(\varepsilon_{\widetilde{n}})$ cannot shrink too fast. Having relabeled the indices appropriately, we can take limits along the diagonal to obtain the desired result.

That said, let's proceed with the formal proof.

Part 1. Consider

$$
\begin{aligned}
&|\widehat{f}_{\pi_{12},\pi_{22}}(p_1,p_2;n,M,\delta) - f_{\pi_{12},\pi_{22}}(p_1,p_2)| \\
&= \left| \mathrm{Re}\left[\frac{1}{(2\pi)^2}\int_{[-M,M]^2} \exp(-i[t_1p_1+t_2p_2])\widehat{\phi}_{\pi_{12},\pi_{22}}(t_1,t_2;n,\delta)\ dt_1dt_2\right]\right. \\
&\qquad \left. - \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} \exp(-i[t_1p_1+t_2p_2])\phi_{\pi_{12},\pi_{22}}(t_1,t_2)\ dt_1dt_2 \right| \\
&\leq \frac{1}{(2\pi)^2}\int_{[-M,M]^2\setminus(-\delta,\delta)^2} |\exp(-i[t_1p_1+t_2p_2])| \cdot |\widehat{\phi}_{\pi_{12},\pi_{22}}(t_1,t_2;n) - \phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\ dt_1dt_2 \\
&\qquad + \frac{1}{(2\pi)^2}\int_{(-\delta,\delta)^2} |\exp(-i[t_1p_1+t_2p_2])| \cdot |1 - \phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\ dt_1dt_2 \\
&\qquad + \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2\setminus[-M,M]^2} |\exp(-i[t_1p_1+t_2p_2])| \cdot |\phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\ dt_1dt_2 \\
&= \frac{1}{(2\pi)^2}\int_{[-M,M]^2\setminus(-\delta,\delta)^2} 1 \cdot |\widehat{\phi}_{\pi_{12},\pi_{22}}(t_1,t_2;n) - \phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\ dt_1dt_2 \\
&\qquad + \frac{1}{(2\pi)^2}\int_{(-\delta,\delta)^2} 1 \cdot |1 - \phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\ dt_1dt_2 \\
&\qquad + \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2\setminus[-M,M]^2} 1 \cdot |\phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\ dt_1dt_2 \\
&\equiv R_1(n,M,\delta) + R_2(\delta) + R_3(M).
\end{aligned}
$$

The first line follows by definition of our third step estimator. In the final line, p_1 and p_2 no longer appear on the right hand side and hence all convergence on the right hand side is uniform over p_1 and p_2. $R_2(\delta)$ can be made arbitrarily small by picking δ small enough, since $\phi_{\pi_{12},\pi_{22}}(t_1,t_2)$ is uniformly continuous and converges to 1 as $(t_1,t_2) \to (0,0)$. $R_3(M)$ is finite for all M by our assumption that $\phi_{\pi_{12},\pi_{22}}$ is integrable. $R_3(M)$ can thus be made arbitrarily small by choosing M large enough, by the Riemann-Lebesgue lemma.

Finally, consider the first term. We have

$$
\begin{aligned}
&R_1(n,M,\delta)\\
&= \frac{1}{(2\pi)^2}\int_{[-M,M]^2\setminus(-\delta,\delta)^2} |\widehat{\phi}_{\pi_{12},\pi_{22}}(t_1,t_2;n) - \phi_{\pi_{12},\pi_{22}}(t_1,t_2)|\, dt_1 dt_2\\
&= \frac{1}{(2\pi)^2}\int_{[-M,M]^2\setminus(-\delta,\delta)^2} \left| \int_{-B(t_1,t_2)}^{B(t_1,t_2)} \exp(is)[\widehat{f}_{\Pi_2(t_1,t_2)}(s;n) - f_{\Pi_2(t_1,t_2)}(s)]\, ds \right| dt_1 dt_2\\
&\le \frac{1}{(2\pi)^2}\int_{[-M,M]^2\setminus(-\delta,\delta)^2} \left(\int_{-B(t_1,t_2)}^{B(t_1,t_2)} |\exp(is)|\cdot|\widehat{f}_{\Pi_2(t_1,t_2)}(s;n) - f_{\Pi_2(t_1,t_2)}(s)|\, ds \right) dt_1 dt_2\\
&\le \frac{1}{(2\pi)^2}\int_{[-M,M]^2\setminus(-\delta,\delta)^2} \left(\sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s;n) - f_{\Pi_2(t_1,t_2)}(s)| \int_{-B(t_1,t_2)}^{B(t_1,t_2)} ds \right) dt_1 dt_2\\
&\le \frac{[(2M)^2-(2\delta)^2][2\overline{B}(M,\delta)]}{(2\pi)^2} \sup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2}\ \sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s;n) - f_{\Pi_2(t_1,t_2)}(s)|.
\end{aligned}
$$

The second line follows by definition of our second step estimator. In the final line,

$$\overline{B}(M,\delta) = \sup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2} B(t_1,t_2).$$

E2 now shows that the right hand side of the above inequality converges almost surely to zero. Hence we have shown that, for any fixed M and δ, $R_1(n,M,\delta)\to 0$ a.s.

Let us put this all together to conclude part 1. Fix an $\varepsilon>0$. Choose $M(\varepsilon)$ large enough and $\delta(\varepsilon)$ small enough that $R_2(\delta(\varepsilon))\le\varepsilon/3$ and $R_3(M(\varepsilon))\le\varepsilon/3$. Since $R_1(n,M(\varepsilon),\delta(\varepsilon))\to 0$ a.s., there is an $N(\varepsilon,M(\varepsilon),\delta(\varepsilon))$ such that for all $n\ge N(\varepsilon,M(\varepsilon),\delta(\varepsilon))$, $R_1(n,M(\varepsilon),\delta(\varepsilon))\le\varepsilon/3$ a.s. Moreover, from here on I let $N(\varepsilon,M(\varepsilon),\delta(\varepsilon))$ denote the smallest such number.

Hence for all $n\ge N(\varepsilon,M(\varepsilon),\delta(\varepsilon))$,

$$\sup_{(p_1,p_2)\in\mathbb{R}^2} |\widehat{f}_{\pi_{12},\pi_{22}}(p_1,p_2;n,M(\varepsilon),\delta(\varepsilon)) - f_{\pi_{12},\pi_{22}}(p_1,p_2)| \le \varepsilon \qquad a.s.$$

Part 2. For notational clarity, define

$$A_n(M,\delta) = \sup_{(p_1,p_2)\in\mathbb{R}^2} |\widehat{f}_{\pi_{12},\pi_{22}}(p_1,p_2;n,M,\delta) - f_{\pi_{12},\pi_{22}}(p_1,p_2)|.$$

Thus, using this notation, from part 1 we showed that for any $\varepsilon>0$, there exists an $M(\varepsilon)$, a $\delta(\varepsilon)$, and an $N(\varepsilon,M(\varepsilon),\delta(\varepsilon))$ such that for all $n\ge N(\varepsilon,M(\varepsilon),\delta(\varepsilon))$,

$$A_n(M(\varepsilon),\delta(\varepsilon))\le\varepsilon \qquad a.s.$$

In this part we want to show that there exists a sequence (M_n, δ_n) such that

$$A_n(M_n, \delta_n) \to 0 \qquad a.s.$$

as $n \to \infty$.

We begin with an arbitrarily chosen decreasing sequence of positive numbers ε^n such that $\varepsilon^n \searrow 0$ as $n \to \infty$. (Superscripts here refer to indices, not powers.) I use this sequence to construct a new sequence ε_n that has the following properties:

1. $n \geq N(\varepsilon_n, M(\varepsilon_n), \delta(\varepsilon_n))$ for all $n \in \mathbb{N}$ (or at least for all sufficiently large n, since the beginning of any sequence is irrelevant for limiting properties).

2. $\varepsilon_n \searrow 0$ as $n \nearrow \infty$.

Here is an algorithm which produces such a sequence.

1. Define
$$\varepsilon_{N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0))} = \varepsilon^0.$$

2. Suppose $N(\varepsilon^1, M(\varepsilon^1), \delta(\varepsilon^1)) > N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0))$ (see step 4 for what to do if this doesn't hold). Then define
$$\varepsilon_{N(\varepsilon^1, M(\varepsilon^1), \delta(\varepsilon^1))} = \varepsilon^1.$$

3. For any integers k such that
$$N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0)) < k < N(\varepsilon^1, M(\varepsilon^1), \delta(\varepsilon^1))$$
define
$$\varepsilon_k = \varepsilon^0.$$

4. And so on. If at any step it happens to be that
$$N(\varepsilon^k, M(\varepsilon^k), \delta(\varepsilon^k)) = N(\varepsilon^{k+1}, M(\varepsilon^{k+1}), \delta(\varepsilon^{k+1}))$$
then proceed to the next $\varepsilon^{k+\ell}$ in the sequence which yields
$$N(\varepsilon^k, M(\varepsilon^k), \delta(\varepsilon^k)) < N(\varepsilon^{k+\ell}, M(\varepsilon^{k+\ell}), \delta(\varepsilon^{k+\ell})).$$

For example, suppose $N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0)) = 3$. Then in step 1 we defined $\varepsilon_3 = \varepsilon^0$. We do not need to define ε_1 and ε_2 because the beginning of a sequence is irrelevant for its limiting properties. Suppose next that $N(\varepsilon^1, M(\varepsilon^1), \delta(\varepsilon^1)) = 6$. Then in step 2 we defined $\varepsilon_6 = \varepsilon^1$. Note $\varepsilon_6 \leq \varepsilon_3$ since $\varepsilon^1 \leq \varepsilon^0$. We skipped a few integers, so in step 3 we define $\varepsilon_4 = \varepsilon^0$ and $\varepsilon_5 = \varepsilon^0$. This ensures that $\varepsilon_6 \leq \varepsilon_5$ and $\varepsilon_6 \leq \varepsilon_4$. Suppose next that $N(\varepsilon^2, M(\varepsilon^2), \delta(\varepsilon^2)) = 6$. We have already defined ε_6 so we proceed to look at $N(\varepsilon^3, M(\varepsilon^3), \delta(\varepsilon^3))$. Suppose this equals 8. Then we let $\varepsilon_8 = \varepsilon^3$. We then fill the gap by letting $\varepsilon_7 = \varepsilon^3$. And so on.

Next I show that $\overline{N}(\varepsilon) \equiv N(\varepsilon, M(\varepsilon), \delta(\varepsilon))$ weakly increases towards ∞ as ε decreases to zero. This ensures that the algorithm actually defines a complete infinite sequence.

- Proof of this step: Let $\varepsilon \leq \widetilde{\varepsilon}$. We want to show that $\overline{N}(\varepsilon) \geq \overline{N}(\widetilde{\varepsilon})$. By definition of $\overline{N}$ in part 1, we have that $n \geq \overline{N}(\varepsilon)$ implies

  $$R_1(n, M(\varepsilon), \delta(\varepsilon)) \leq \varepsilon/3 \qquad a.s.$$

  But since $\varepsilon \leq \widetilde{\varepsilon}$, we have that $n \geq \overline{N}(\varepsilon)$ also implies

  $$R_1(n, M(\varepsilon), \delta(\varepsilon)) \leq \widetilde{\varepsilon}/3 \qquad a.s.$$

  Next, note that

  - $\delta(\cdot)$ is weakly increasing; as $\varepsilon \searrow 0$, $\delta(\varepsilon) \searrow 0$. Hence $\delta(\varepsilon) \leq \delta(\widetilde{\varepsilon})$.
  - $M(\cdot)$ is weakly decreasing; as $\varepsilon \searrow 0$, $M(\varepsilon) \nearrow \infty$. Hence $M(\varepsilon) \geq M(\widetilde{\varepsilon})$.
  - $R_1(n, M, \delta)$ is weakly increasing as M gets larger and is weakly increasing as δ gets smaller. This follows by because larger M and smaller δ only affect R_1 by increasing the domain of integration, and the integrand is nonnegative.

  Hence

  $$R_1(n, M(\varepsilon), \delta(\varepsilon)) \geq R_1(n, M(\widetilde{\varepsilon}), \delta(\widetilde{\varepsilon})).$$

  Combining this with our derivations above yields that $n \geq \overline{N}(\varepsilon)$ implies

  $$R_1(n, M(\widetilde{\varepsilon}), \delta(\widetilde{\varepsilon})) \leq \widetilde{\varepsilon}/3 \qquad a.s.$$

  Finally, since $\overline{N}(\widetilde{\varepsilon})$ was defined to be the *smallest* such number that this holds, we must have $\overline{N}(\widetilde{\varepsilon}) \leq \overline{N}(\varepsilon)$, as desired.

- That shows that $\overline{N}(\varepsilon)$ is weakly increasing as ε decreases to zero. It is possible that $\overline{N}(\varepsilon) \nearrow Z$ as $\varepsilon \searrow 0$, where Z is some fixed finite integer. In this case, however, we have that for all $n \geq Z$

  $$R_1(n, M(\varepsilon), \delta(\varepsilon)) \leq \varepsilon/3 \qquad a.s.$$

  for all $\varepsilon > 0$. This implies

  $$R_1(n, M(\varepsilon), \delta(\varepsilon)) = 0 \qquad a.s.$$

  for all $n \geq Z$, since $R_1 \geq 0$. That is, we obtain convergence at a finite sample size, rather than asymptotically. Consequently, for any fixed $n \geq Z$ we can make $A_n(M, \delta)$ arbitrarily small by picking M and δ appropriately. Thus the case where $\overline{N}(\varepsilon) \nearrow Z$ is a trivial case. Instead, I focus on the nontrivial case where $\overline{N}(\varepsilon) \nearrow \infty$ as $\varepsilon \searrow 0$.

Next I show that this newly constructed sequence ε_n has the two properties we want.

1. The property $n \geq N(\varepsilon_n, M(\varepsilon_n), \delta(\varepsilon_n))$ holds by construction. For example, return to the example of our algorithm. Here $N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0)) = 3$. Then we defined $\varepsilon_3 = \varepsilon^0$. Hence

   $$N(\varepsilon_3, M(\varepsilon_3), \delta(\varepsilon_3)) = N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0)) = 3.$$

   Thus for $n = 3$,

   $$n \geq N(\varepsilon_n, M(\varepsilon_n), \delta(\varepsilon_n))$$

   is $3 \geq 3$ and hence it actually holds with equality. Next consider $n = 4$. Here we defined $\varepsilon_4 = \varepsilon^0$. Hence

   $$N(\varepsilon_4, M(\varepsilon_4), \delta(\varepsilon_4)) = N(\varepsilon^0, M(\varepsilon^0), \delta(\varepsilon^0)) = 3.$$

   Thus for $n = 4$,

   $$n \geq N(\varepsilon_n, M(\varepsilon_n), \delta(\varepsilon_n))$$

   is $4 \geq 3$ and hence it holds with a strict inequality. In general, there is a subsequence of $\{\varepsilon_n\}$ such that the inequality holds with equality. And for all other elements the inequality holds as a strict inequality.

2. Our constructed sequence ε_n decreases to zero because (a) ε^n decreases to zero and (b) $\overline{N}(\varepsilon)$ is weakly increasing towards ∞ as ε decreases to zero. (b) ensures that there is a bijection between a subsequence of $\{\varepsilon_n\}$ and the sequence $\{\varepsilon^n\}$. Hence, by (a), zero is a limit point of $\{\varepsilon_n\}$. But since ε_n is weakly decreasing and is nonnegative, the overall sequence must converge to zero.

Intuitively, this algorithm takes an arbitrary sequence ε^n that decreases towards zero and stretches it out (equivalently, slows it down) by relabeling the indices, so that it converges to zero slow enough that $n \geq N(\varepsilon_n, M(\varepsilon_n), \delta(\varepsilon_n))$ always holds.

A few other remarks:

- This algorithm is not helpful in practice since the function $N(\cdot, \cdot, \cdot)$ is unknown. But our goal here is merely to prove the existence of a sequence of valid tuning parameters. For this goal, the above algorithm suffices.

- If we didn't require that $\varepsilon_n \searrow 0$ as $n \nearrow \infty$, then requirement (1) for our sequence could always be satisfied by letting ε_n be arbitrarily large for all n (this follows since $\overline{N}(\varepsilon)$ is weakly decreasing as ε gets large).

Next, we use this special sequence ε_n to define our tuning parameter sequences, as follows:

$$M_n^* = M(\varepsilon_n) \qquad \text{and} \qquad \delta_n^* = \delta(\varepsilon_n).$$

I conclude part 2 by showing that this sequence of tuning parameters works.

First consider a single fixed $\widetilde{n}$. By part 1, we have

$$A_n(M(\varepsilon_{\widetilde{n}}), \delta(\varepsilon_{\widetilde{n}})) \leq \varepsilon_{\widetilde{n}} \qquad a.s.$$

for all $n \geq N(\varepsilon_{\widetilde{n}}, M(\varepsilon_{\widetilde{n}}), \delta(\varepsilon_{\widetilde{n}}))$. In particular, this holds for $n = \widetilde{n}$, by property (1) of our choice of $\varepsilon_{\widetilde{n}}$. Hence

$$A_{\widetilde{n}}(M(\varepsilon_{\widetilde{n}}), \delta(\varepsilon_{\widetilde{n}})) \leq \varepsilon_{\widetilde{n}}. \qquad a.s.$$

This holds for every $\widetilde{n}$. Next take limits of both sides as $\widetilde{n} \to 0$. On the right hand side we have $\varepsilon_{\widetilde{n}} \to 0$ by property (2) of our choice of $\varepsilon_{\widetilde{n}}$. The left hand side is nonnegative. Hence

$$A_{\widetilde{n}}(M(\varepsilon_{\widetilde{n}}), \delta(\varepsilon_{\widetilde{n}})) \to 0 \qquad a.s.$$

as $\widetilde{n} \to \infty$. That is, we have shown that

$$\sup_{(p_1,p_2)\in\mathbb{R}^2} |\widehat{f}_{\pi_{12},\pi_{22}}(p_1, p_2; n, M_n^*, \delta_n^*) - f_{\pi_{12},\pi_{22}}(p_1, p_2)| \to 0 \qquad a.s.$$

as $n \to \infty$ (where replacing $\widetilde{n}$ with n is just a matter of notation).

Part 3. Next consider the fourth step estimator:

$$\begin{aligned}
|\widehat{f}_{\gamma_2}(z) - f_{\gamma_2}(z)| &= \left| \int_{\mathbb{R}} |v| \widehat{f}_{\pi_{12},\pi_{22}}(v, zv)\, dv - \int_{\mathbb{R}} |v| f_{\pi_{12},\pi_{22}}(v, zv)\, dv \right| \\
&\leq \int_{\mathbb{R}} |v| \big| \widehat{f}_{\pi_{12},\pi_{22}}(v, zv) - f_{\pi_{12},\pi_{22}}(v, zv) \big|\, dv \\
&= \int_{[-D,D]^2} |v| \big| \widehat{f}_{\pi_{12},\pi_{22}}(v, zv) - f_{\pi_{12},\pi_{22}}(v, zv) \big|\, dv \\
&\leq \sup_{(p_1,p_2)\in\mathbb{R}^2} |\widehat{f}_{\pi_{12},\pi_{22}}(p_1, p_2) - f_{\pi_{12},\pi_{22}}(p_1, p_2)| \int_{[-D,D]^2} |v|\, dv.
\end{aligned}$$

In the third line we let $\text{supp}(\pi_{12}, \pi_{22}) \subseteq [-D, D]^2$ where $0 < D < \infty$, which is possible by our compact support assumption E1.2. The integral in the fourth line is finite. Hence sup-norm consistency of $\widehat{f}_{\pi_{12},\pi_{22}}$ implies sup-norm consistency of our main estimator of interest:

$$\sup_{z\in\mathbb{R}} |\widehat{f}_{\gamma_2}(z) - f_{\gamma_2}(z)| \to 0 \qquad a.s.$$

as $n \to \infty$, as desired. □

Next I show that a slightly stronger version of assumption E2 leads to a simpler consistency proof.

Assumption E2′ We have

$$\sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s; n) - f_{\Pi_2(t_1,t_2)}(s)| \leq C(t_1, t_2) \cdot a_n \qquad a.s.$$

where $a_n \to 0$ as $n \to \infty$ and for any fixed $M > \delta > 0$,

$$\sup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2} C(t_1,t_2) < \infty \qquad a.s.$$

As with assumption E2, this is a high-level assumption on the first stage estimator. E2′ has two main parts. First, it states that the uniform convergence rate a_n does not depend on (t_1, t_2). (t_1, t_2) are parameters that affect the true density we are estimating. Hence E2′ states that the uniform rate of convergence does not vary over the class of densities under consideration. Since this typically holds for kernel density estimators, we might expect that it will also hold for estimators like the inverse Radon transform estimator. Indeed, Bissantz, Holzmann, and Proksch (2014) prove a similar result for a closely related estimator in their remark 3 on page 92.

Second, E2′ states that the constant term $C(t_1, t_2)$ uniformly bounded over a specific compact set of (t_1, t_2)'s not containing zero. This assumption also typically holds for kernel density estimators. For example, see theorem B of Silverman (1978), or the proof of theorem 1.4 of Li and Racine (2007), where the constant depends on objects like

$$C_1(t_1,t_2) = \sup_{s\in[-B(t_1,t_2),B(t_t,t_2)]} |f_{\Pi_2(t_1,t_2)}(s)|.$$

Continuity of $f_{\Pi_2(t_1,t_2)}$ in (t_1, t_2) implies (lemma S3 below) implies that $C_1(t_1, t_2)$ is uniformly bounded on $[-M, M]^2 \setminus (-\delta, \delta)^2$. Similar results can likely also be shown for suprema over higher order derivatives of $f_{\Pi_2(t_1,t_2)}$, which may also appear in the constant term $C(t_1, t_2)$. Note that here $\delta > 0$ will be crucial to prevent the true density from becoming unbounded as either t_1 or t_2 goes to zero. The Bissantz et al. (2014) result is uniform over a Sobolev class of densities which may also be helpful in showing E2′ for the first stage estimators of interest.

E2′ implies E2, but it allows us to provide a simpler proof of theorem S2, as follows. A key observation is that M and δ do not enter the first stage estimator.

Proof of theorem S2 under E2′. Here I show how to modify the previous proof appropriately. From part 1, we have

$$\begin{aligned} R_1(n,M,\delta) &\leq \frac{[(2M)^2-(2\delta)^2][2\overline{B}(M,\delta)]}{(2\pi)^2} \\ &\qquad \cdot \sup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2} \; \sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s;n) - f_{\Pi_2(t_1,t_2)}(s)| \\ &\leq \frac{[(2M)^2-(2\delta)^2][2\overline{B}(M,\delta)]}{(2\pi)^2} \sup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2} C(t_1,t_2)a_n \qquad a.s. \\ &\leq a_n \cdot \frac{[(2M)^2-(2\delta)^2][2\overline{B}(M,\delta)]}{(2\pi)^2} \overline{C}(M,\delta) \qquad a.s. \end{aligned}$$

The second and third lines follow by E2′, where $\overline{C}(M,\delta) < \infty$. Thus we have

$$\begin{aligned}\sup_{(p_1,p_2)\in\mathbb{R}^2} &|\widehat{f}_{\pi_{12},\pi_{22}}(p_1,p_2;n,M,\delta) - f_{\pi_{12},\pi_{22}}(p_1,p_2)| \\ &\qquad\qquad \leq R_1(n,M,\delta) + R_2(\delta) + R_3(M) \\ &\qquad\qquad \leq a_n \cdot \frac{1}{(2\pi)^2} 8\overline{B}(M,\delta)\overline{C}(M,\delta)(M^2-\delta^2) + R_2(\delta) + R_3(M) \qquad a.s.\end{aligned}$$

Now we need to pick $M_n \to \infty$ and $\delta_n \to 0$ so that the second two terms converge to zero. But we need M_n to grow slow enough and δ_n to shrink slow enough that the first term also converges to zero. Specifically, we need

$$\frac{1}{(2\pi)^2} 8\overline{B}(M_n,\delta_n)\overline{C}(M_n,\delta_n)(M_n^2-\delta_n^2) = o(1/a_n).$$

This concludes the modified proof of part 1. We no longer need part 2, and part 3 is exactly as before. □

Next I consider verification of E2 for specific estimators. The literature on estimating distributions of random coefficients in single equation models is quite unexplored. Very few sup-norm convergence results exist. Beran and Hall (1992) briefly sketch an argument for almost sure sup-norm consistency of their estimator on page 1974. In their remark 3 on page 92, Bissantz et al. (2014) prove sup-norm consistency in probability for an inverse Radon transform estimator similar to that of Hoderlein et al. (2010), in a Gaussian model. Such results for the other estimators remain to be shown, however.

The papers of Beran et al. (1996), Hoderlein et al. (2010), and Beran and Millar (1994) do not provide sup-norm convergence results. Beran et al. (1996) prove density-weighted L_2 consistency of their estimator, while Hoderlein et al. (2010) prove unweighted L_2 consistency of their estimator. Beran and Millar (1994) do not actually estimate pdfs; they only estimate cdfs and prove consistency in the weak convergence topology.

Since sup-norm convergence is still an open question in single equation models, here I only sketch a result that shows how the extra supremum over (t_1,t_2) in E2 can be handled similarly to the supremum over s. I omit the specific rate constraints, however.

Define

$$\widehat{Q}(t_1,t_2) = \sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s) - f_{\Pi_2(t_1,t_2)}(s)|$$

and

$$\mathcal{D} = [-M,M]^2 \setminus (-\delta,\delta)^2.$$

We want to show that for any fixed M and δ,

$$\sup_{(t_1,t_2)\in\mathcal{D}} \widehat{Q}(t_1,t_2) \to 0 \qquad a.s.$$

as $n \to \infty$. Since $\mathcal{D}$ is compact (it's a square donut), we can cover it by a finite number of balls $\mathcal{D}_1, \ldots, \mathcal{D}_{K_n}$ with centers (t_{1k}, t_{2k}) and radii rad_n. Then

$$\sup_{(t_1,t_2)\in\mathcal{D}} \widehat{Q}(t_1,t_2) \leq \max_{1\leq k\leq K_n} \sup_{(t_1,t_2)\in\mathcal{D}\cap\mathcal{D}_{K_n}} |\widehat{Q}(t_1,t_2) - \widehat{Q}(t_{1k},t_{2k})| + \max_{1\leq k\leq K_n} \widehat{Q}(t_{1k},t_{2k}). \qquad (*)$$

For fixed K_n, the second term converges so long as a usual sup-norm convergence result holds. For $K_n \to \infty$, however, we need to ensure that K_n cannot grow too fast in order for this second term to converge to zero. The formal analysis of this case can likely be adapted from a proof of sup-norm convergence for fixed (t_1, t_2), which are currently unavailable in the literature for the estimators and assumptions of interest here.

Consider the first term. We have

$$\begin{aligned}\widehat{Q}(t_1,t_2) - \widehat{Q}(k_1,k_2) = &\sup_{s\in[-B(t_1,t_2),B(t_1,t_2)]} |\widehat{f}_{\Pi_2(t_1,t_2)}(s) - f_{\Pi_2(t_1,t_2)}(s)| \\ &- \sup_{s\in[-B(k_1,k_2),B(k_1,k_2)]} |\widehat{f}_{\Pi_2(k_1,k_2)}(s) - f_{\Pi_2(k_1,k_2)}(s)|.\end{aligned}$$

Since these suprema are of continuous functions over compact sets, they are achieved. Let $\tilde{s}$ denote the point at which the first supremum on the right hand side is achieved ($\tilde{s}$ depends on t_1, t_2). Then we have

$$\begin{aligned}\widehat{Q}(t_1,t_2) - \widehat{Q}(t_{1k},t_{2k}) &= |\widehat{f}_{\Pi_2(t_1,t_2)}(\tilde{s}) - f_{\Pi_2(t_1,t_2)}(\tilde{s})| \\ &\qquad - \sup_{s\in[-B(t_{1k},t_{2k}),B(t_{1k},t_{2k})]} |\widehat{f}_{\Pi_2(t_{1k},t_{2k})}(s) - f_{\Pi_2(t_{1k},t_{2k})}(s)| \\ &\leq |\widehat{f}_{\Pi_2(t_1,t_2)}(\tilde{s}) - f_{\Pi_2(t_1,t_2)}(\tilde{s})| - |\widehat{f}_{\Pi_2(t_{1k},t_{2k})}(\tilde{s}) - f_{\Pi_2(t_{1k},t_{2k})}(\tilde{s})|.\end{aligned}$$

The second line follows by definition of the supremum. Taking absolute values of both sides,

$$\begin{aligned}|\widehat{Q}(t_1,t_2) - \widehat{Q}(t_{1k},t_{2k})| &\leq \Big| |\widehat{f}_{\Pi_2(t_1,t_2)}(\tilde{s}) - f_{\Pi_2(t_1,t_2)}(\tilde{s})| - |\widehat{f}_{\Pi_2(t_{1k},t_{2k})}(\tilde{s}) - f_{\Pi_2(t_{1k},t_{2k})}(\tilde{s})| \Big| \\ &\leq \Big| \big(\widehat{f}_{\Pi_2(t_1,t_2)}(\tilde{s}) - \widehat{f}_{\Pi_2(t_{1k},t_{2k})}(\tilde{s})\big) - \big(f_{\Pi_2(t_1,t_2)}(\tilde{s}) - f_{\Pi_2(t_{1k},t_{2k})}(\tilde{s})\big) \Big| \\ &\leq |\widehat{f}_{\Pi_2(t_1,t_2)}(\tilde{s}) - \widehat{f}_{\Pi_2(t_{1k},t_{2k})}(\tilde{s})| + |f_{\Pi_2(t_1,t_2)}(\tilde{s}) - f_{\Pi_2(t_{1k},t_{2k})}(\tilde{s})| \\ &\leq A_n g(h)\|(t_1,t_2) - (t_{1k},t_{2k})\| + B\|(t_1,t_2) - (t_{1k},t_{2k})\| \\ &= (A_n g(h) + B)\|(t_1,t_2) - (t_{1k},t_{2k})\| \\ &\leq (A_n g(h) + B)\text{rad}_n.\end{aligned}$$

Here $A_n = O_p(1)$ and $g(h)$ is a function of the first stage estimator bandwidth h such that $g(h) \to \infty$ as $h \to 0$. The second line follows by the reverse triangle inequality, while the third line follows by the ordinary triangle inequality. The fourth line follows by lemma S3 and, for the estimator of Hoderlein et al. (2010), by lemma S2. Note that $\tilde{s}$ drops out at line four.

Here we see that as long as the radius of our covering balls rad_n shrinks fast enough (equivalently,

K_n grows fast enough) that $g(h_n)\text{rad}_n \to 0$ then

$$\max_{1 \leq k \leq K_n} \sup_{(t_1,t_2) \in \mathcal{D} \cap \mathcal{D}_{K_n}} |\widehat{Q}(t_1,t_2) - \widehat{Q}(t_{1k},t_{2k})| \to 0 \qquad a.s.$$

Thus we see that, as usual when proving uniform convergence, the rate K_n must balance each of the two terms in equation $(*)$ so that both terms converge to zero.

I conclude this section with two lemmas used in the above derivations.

Lemma S2. Let $\widehat{f}_{\Pi_2(t_1,t_2)}(s)$ denote the inverse Radon transform estimator of Hoderlein et al. (2010) as applied in step 1 of section 4. r in their paper is essentially a kernel tuning parameter. Let $r = 1$. f_S in their paper is the density of the covariates after scaling them to have unit norm. Assume f_S is bounded away from zero (their assumption 4). Suppose the estimator $\widehat{f}_S$ converges uniformly almost surely to f_S (almost sure version of their assumption 3). Assume the means of $\tilde{Y}_1$ and $\tilde{Y}_2$ exist, where these are the original outcome variables divided by the norm of the covariate vector. Let h denote the bandwidth parameter. Then there is a sequence of random variables $A_n = O_p(1)$ and a function $g(h) = 1/h^4$ such that

$$|\widehat{f}_{\Pi_2(t_1,t_2)}(b_2) - \widehat{f}_{\Pi_2(k_1,k_2)}(b_2)| \leq A_n g(h) \|(t_1,t_2) - (k_1,k_2)\|$$

for any $(t_1,t_2),(k_1,k_2) \in \mathbb{R}^2$ and $b_2 \in \mathbb{R}$.

See Hoderlein et al. (2010) for discussion of their assumptions 3 and 4.

Proof of lemma S2. In this proof I freely use the notation of Hoderlein et al. (2010). Let S_i denote the ith covariate scaled to have unit norm. Let

$$U_i = t_1 \tilde{Y}_{1i} + t_2 \tilde{Y}_{2i}$$

denote the outcome variable, where $\tilde{Y}_{1i}$ and $\tilde{Y}_{2i}$ are the original outcome variables divided by the norm of the ith covariate. Then the inverse Radon transform estimator of $f_{\Pi(t_1,t_2)}$ is

$$\widehat{f}_{\Pi(t_1,t_2)}(b) = \frac{1}{n} \sum_{i=1}^{n} \frac{K_{r,h}(S_i' b - U_i)}{\widehat{f}_S(S_i)}$$

where $b = (b_1, b_2, b_3)$ is a 3-dimensional vector, since $\Pi(t_1,t_2)$ is a 3-dimensional random variable. For $r = 1$ and $\dim[\Pi(t_1,t_2)] = 3$ the kernel function is

$$K_{1,h}(a) = \frac{1}{h^3} \mathcal{K}_1 \left(\frac{a}{h} \right)$$

where

$$\mathcal{K}_1(u) = \begin{cases} \dfrac{1}{12} & \text{if } u = 0 \\[2em] \dfrac{1}{8\pi^3}\left(\dfrac{4u\sin(u) - 6 - (u^2-6)\cos(u)}{u^4}\right) & \text{if } u \neq 0. \end{cases}$$

$\mathcal{K}_1$ is Lipschitz with constant $C_{\mathcal{K}}$.

To obtain our estimator of $f_{\Pi_2(t_1,t_2)}$ we numerically integrate over the b_1 and b_3 components:

$$\begin{aligned} \widehat{f}_{\Pi_2(t_1,t_2)}(b_2) &= (2B)^2 \frac{1}{L}\sum_{\ell=1}^{L} \widehat{f}_{\Pi(t_1,t_2)}(b_{1\ell}, b_2, b_{3\ell}) \\ &= (2B)^2 \frac{1}{L}\sum_{\ell=1}^{L} \frac{1}{n}\sum_{i=1}^{n} \frac{K_{r,h}(S_i' b_\ell - U_i)}{\widehat{f}_S(S_i)} \end{aligned}$$

where $b_\ell = (b_{1\ell}, b_2, b_{3\ell})$, and $\{(b_{1\ell}, b_{3\ell})\}_{\ell=1}^{L}$ is a user chosen grid of equidistributed points in the square $[-B,B]^2$ where $B \to \infty$ and $L \to \infty$ as $n \to \infty$. This is just a numerical approximation; B and L can be chosen arbitrarily large for any fixed n, up to computational constraints. In particular, for the purposes of this proof, B can be considered to be fixed, similarly to M.

We have

$$\begin{aligned} &|\widehat{f}_{\Pi_2(t_1,t_2)}(b_2) - \widehat{f}_{\Pi_2(k_1,k_2)}(b_2)| \\ &= \left|(2B)^2 \frac{1}{L}\sum_{\ell=1}^{L} \frac{1}{n}\sum_{i=1}^{n} \left(\frac{K_{1,h}(S_i' b_\ell - [t_1 \tilde{Y}_{1i} + t_2 \tilde{Y}_{2i}])}{\widehat{f}_S(S_i)} - \frac{K_{1,h}(S_i' b_\ell - [k_1 \tilde{Y}_{1i} + k_2 \tilde{Y}_{2i}])}{\widehat{f}_S(S_i)} \right)\right| \\ &\le \frac{(2B)^2}{C_1 L}\sum_{\ell=1}^{L} \frac{1}{n}\sum_{i=1}^{n} \left| K_{1,h}(S_i' b_\ell - [t_1 \tilde{Y}_{1i} + t_2 \tilde{Y}_{2i}]) - K_{1,h}(S_i' b_\ell - [k_1 \tilde{Y}_{1i} + k_2 \tilde{Y}_{2i}]) \right| \\ &= \frac{(2B)^2}{C_1 L}\sum_{\ell=1}^{L} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^3}\left| \mathcal{K}_1\left(\frac{S_i' b_\ell - [t_1 \tilde{Y}_{1i} + t_2 \tilde{Y}_{2i}]}{h}\right) - \mathcal{K}_1\left(\frac{S_i' b_\ell - [k_1 \tilde{Y}_{1i} + k_2 \tilde{Y}_{2i}]}{h}\right) \right| \\ &\le \frac{(2B)^2}{C_1 L}\sum_{\ell=1}^{L} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^3}\frac{C_{\mathcal{K}}}{h} |(S_i' b_\ell - [t_1 \tilde{Y}_{1i} + t_2 \tilde{Y}_{2i}]) - (S_i' b_\ell - [k_1 \tilde{Y}_{1i} + k_2 \tilde{Y}_{2i}])| \\ &\le \frac{C_{\mathcal{K}}(2B)^2}{C_1 L}\sum_{\ell=1}^{L} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^3}\frac{1}{h} |\tilde{Y}_{1i}(k_1 - t_1) + \tilde{Y}_{2i}(k_2 - t_2)| \\ &= \frac{C_{\mathcal{K}}(2B)^2}{C_1} \frac{1}{h^4} \frac{1}{n}\sum_{i=1}^{n} |\tilde{Y}_{1i}(k_1 - t_1) + \tilde{Y}_{2i}(k_2 - t_2)| \\ &\le \frac{C_{\mathcal{K}}(2B)^2}{C_1} \frac{1}{h^4} \left[\left(\frac{1}{n}\sum_{i=1}^{n} |\tilde{Y}_{1i}|\right) |k_1 - t_1| + \left(\frac{1}{n}\sum_{i=1}^{n} |\tilde{Y}_{2i}|\right) |k_2 - t_2| \right] \\ &\le A_n g(h) \|(t_1,t_2) - (k_1,k_2)\|. \end{aligned}$$

The second line holds for all n sufficiently large, almost surely. This follows by almost sure uniform convergence of $\widehat{f}_S$ and since f_S is bounded away from zero by C_1. The fourth line follows since $\mathcal{K}_1$ is Lipschitz. The constant A_n is $O_p(1)$ by the law of large numbers, and $g(h) = 1/h^4$. □

The following lemma verifies that the true density of the linear combination $t_1\pi_{12}+t_2\pi_{22}$ satisfies a Lipschitz condition as used in the sketch above.

Lemma S3. Assume

1. $f_{\pi_{12},\pi_{22}}$ is continuously differentiable.

2. $f_{\pi_{12},\pi_{22}}$ has compact support.

Then for any fixed $M > \delta > 0$ there is a constant B such that for any $(t_1, t_2), (k_1, k_2) \in [-M, M]^2 \setminus (-\delta, \delta)^2$,

$$|f_{\Pi_2(t_1,t_2)}(s) - f_{\Pi_2(k_1,k_2)}(s)| \leq B\|(t_1, t_2) - (k_1, k_2)\|$$

for any $s \in \text{supp}[\Pi_2(t_1, t_2)] \cup \text{supp}[\Pi_2(k_1, k_2)]$.

Proof of lemma S3. $f_{\Pi_2(t_1,t_2)}$ is the density of $t_1\pi_{12} + t_2\pi_{22}$. Its cdf, for $t_1 > 0$, is

$$\begin{aligned} F_{\Pi_2(t_1,t_2)}(s) &= \mathbb{P}(t_1\pi_{12} + t_2\pi_{22} \leq s) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{s-t_2p_{22}}{t_1}} f_{\pi_{12},\pi_{22}}(p_{12}, p_{22})\, dp_{12}\, dp_{22}. \end{aligned}$$

Here and below this integral is completely determined by its values on a compact subset of the limits of integration, by our compact support assumption. Moreover, since $f_{\pi_{12},\pi_{22}}$ is continuous and compactly supported, it is bounded. Hence we can exchange integrals and derivatives by Lebesgue's dominated convergence theorem.

By Leibniz' rule,

$$\frac{\partial}{\partial s}\left[\int_{-\infty}^{\frac{s-t_2p_{22}}{t_1}} f_{\pi_{12},\pi_{22}}(p_{12}, p_{22})\, dp_{12}\right] = f_{\pi_{12},\pi_{22}}\left(\frac{s - t_2p_{22}}{t_1}, p_{22}\right)\frac{1}{t_1}.$$

Hence

$$f_{\Pi_2(t_1,t_2)}(s) = \int_{-\infty}^{\infty} f_{\pi_{12},\pi_{22}}\left(\frac{s - t_2p_{22}}{t_1}, p_{22}\right)\frac{1}{t_1}\, dp_{22}.$$

Thus, again using Leibniz' rule,

$$\frac{\partial}{\partial t_1} f_{\Pi_2(t_1,t_2)}(s) = \int_{-\infty}^{\infty}\left[\frac{1}{t_1^2}\left(\frac{\partial f_{\pi_{12},\pi_{22}}}{\partial \pi_{12}}\left(\frac{s - t_2p_{22}}{t_1}, p_{22}\right) - f_{\pi_{12},\pi_{22}}\left(\frac{s - t_2p_{22}}{t_1}, p_{22}\right)\right)\right] dp_{22}.$$

Since $f_{\pi_{12},\pi_{22}}$ is continuously differentiable, this derivative is also continuous in (t_1, t_2, s), over the set $[-M, M]^2 \setminus (-\delta, \delta^2) \times \mathcal{S}$, where

$$\mathcal{S} = \bigcup_{(t_1,t_2)\in[-M,M]^2\setminus(-\delta,\delta)^2} \text{supp}[\Pi_2(t_1, t_2)].$$

A similar argument holds for the partial derivative of $f_{\Pi_2(t_1,t_2)}(s)$ with respect to t_2. Likewise, a similar argument for both partial derivatives holds for $t_1 < 0$.

Fix a value of s. Then for any $(t_1, t_2), (k_1, k_2) \in [-M, M]^2 \setminus (-\delta, \delta)^2$, the mean value theorem and the Cauchy-Schwarz inequality imply that there exists a $\lambda \in (0, 1)$ such that

$$|f_{\Pi_2(t_1,t_2)}(s) - f_{\Pi_2(k_1,k_2)}(s)| \leq |\nabla f_{\Pi_2((1-\lambda)(t_1,t_2)+\lambda(k_1,k_2))}(s)| \cdot \|(t_1, t_2) - (k_1, k_2)\|$$

where $\nabla f_{\Pi_2(t_1,t_2)}(s)$ denotes the gradient with respect to (t_1, t_2). Let

$$B = \sup_{(t_1,t_2,s)\in[-M,M]^2\setminus(-\delta,\delta^2)\times\mathcal{S}} |\nabla f_{\Pi_2((1-\lambda)(t_1,t_2)+\lambda(k_1,k_2))}(s)|.$$

B is finite by continuity of the gradient in (t_1, t_2, s) and since we are taking the supremum over a compact set. Hence the desired Lipschitz condition holds. □

References

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

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

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

Berman, A. and R. J. Plemmons (1979): *Nonnegative matrices in the mathematical sciences*, Academic Press.

Bissantz, N., H. Holzmann, and K. Proksch (2014): "Confidence regions for images observed under the Radon transform," *Journal of Multivariate Analysis*, 128, 86–107.

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

Bramoullé, Y. and R. Kranton (2016): "Games played on networks," *The Oxford Handbook of the Economics of Networks.*

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

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

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

Li, Q. and J. S. Racine (2007): *Nonparametric econometrics: theory and practice*, Princeton University Press.

Liflyand, E., S. Samko, and R. Trigub (2012): "The Wiener algebra of absolutely convergent Fourier integrals: An overview," *Analysis and Mathematical Physics*, 2, 1–68.

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

Pinsky, M. A. (2002): *Introduction to Fourier analysis and wavelets*, vol. 102, American Mathematical Soc.

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

Scott, D. W. (2015): *Multivariate density estimation: Theory, practice, and visualization*, John Wiley & Sons.

Silverman, B. W. (1978): "Weak and strong uniform consistency of the kernel estimate of a density and its derivatives," *The Annals of Statistics*, 6, 177–184.