I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18091

Bivariate Distribution Regression; Theory, Estimation and an Application to Intergenerational Mobility

Victor Chernozhukov
Iván Fernández-Val
Jonas Meier
Aico Van Vuuren
Francis Vella

AUGUST 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 18091

Bivariate Distribution Regression; Theory, Estimation and an Application to Intergenerational Mobility

Victor Chernozhukov
MIT

Iván Fernández-Val
Boston University

Jonas Meier
Swiss National Bank

Aico Van Vuuren
University of Groningen and IZA

Francis Vella
Georgetown University and IZA

AUGUST 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Bivariate Distribution Regression; Theory, Estimation and an Application to Intergenerational Mobility*

We employ distribution regression (DR) to estimate the joint distribution of two outcome variables conditional on chosen covariates. While Bivariate Distribution Regression (BDR) is useful in a variety of settings, it is particularly valuable when some dependence between the outcomes persists after accounting for the impact of the covariates. Our analysis relies on a result from Chernozhukov et al. (2018) which shows that any conditional joint distribution has a local Gaussian representation. We describe how BDR can be implemented and present some associated functionals of interest. As modeling the unexplained dependence is a key feature of BDR, we focus on functionals related to this dependence. We decompose the difference between the joint distributions for different groups into composition, marginal and sorting effects. We provide a similar decomposition for the transition matrices which describe how location in the distribution in one of the outcomes is associated with location in the other. Our theoretical contributions are the derivation of the properties of these estimated functionals and appropriate procedures for inference. Our empirical illustration focuses on intergenerational mobility. Using the Panel Survey of Income Dynamics data, we model the joint distribution of parents' and children's earnings. By comparing the observed distribution with constructed counterfactuals, we isolate the impact of observable and unobservable factors on the observed joint distribution. We also evaluate the forces responsible for the difference between the transition matrices of sons' and daughters'.

JEL Classification: C14, C21

Keywords: bivariate distribution regression, joint distribution, local Gaussian correlation, decomposition, intergenerational mobility

Corresponding author:
Francis Vella
Economics Department
Georgetown University
37th and O Streets, NW
Washington, DC 20057
USA

E-mail: fgv@georgetown.edu

* Aico van Vuuren received financial support from Vetenskapsr° adet, registration number 2020-02423.

1. Introduction

Although distribution regression (DR) was introduced in the 1970s by Williams and Grizzle (1972), and revisited by Foresi and Peracchi (1995) in the 1990s, it has only relatively recently been more actively employed in empirical economic investigations. For example, inference for DR with a continuous outcome variable was only developed in Chernozhukov et al. (2013) and for discrete, mixed discrete or continuous outcomes in Chernozhukov et al. (2020b). These papers have resulted in an examination of the economically interesting functionals which can be obtained from an estimated distribution of an economic variable conditional on a vector of exogenous variables. This has resulted in the use of DR in a range of different models and settings. These include, for example, models with endogeneity (Chernozhukov et al., 2020a), sample selection (Chernozhukov et al., 2018; Fernández-Val et al., 2024a), dynamic panels (Fernandez-Val et al., 2023), and most recently difference-in-difference estimation (Fernández-Val et al., 2024b).

While these extensions cover a wide class of models they, with the exception of some studies we note below, are restricted to models with univariate outcomes. However, for many empirical questions it is necessary to examine how the joint distribution of two outcomes varies in response to changes in the conditioning variables. For example, in investigating household labor supply responses to changes in tax rates, one should quantify how the hours worked by each of the spouses respond and, perhaps more importantly, how changes in their individual labor supplies are correlated, thereby affecting overall household labor supply. Alternatively, in examining the impact of changes in the minimum wage it would be useful to investigate how both wages and hours of work individually and jointly respond. These examples are just illustrative of the large number of economic questions in which it is important to examine more than one outcome and, perhaps more importantly, how the joint distribution of the outcomes is affected.

The estimation of models with multiple outcomes by DR has been considered in Meier (2020) and Wang et al. (2022). Meier (2020) extends univariate DR to the multivariate setting by using a multivariate indicator function at every location of the distribution. Wang et al. (2022) propose a fast factorization method that also employs univariate DR. However, this transformation to a univariate approach is potentially restrictive as it requires a single outcome variable to be constructed from the multiple outcomes of interest. As DR then employs this constructed outcome, it does not enable each of the individual outcomes to be modeled separately. Moreover, it also bypasses the modeling of the correlation between the unobservables driving each of the outcome variables.In contrast, BDR enables each of the individual outcomes to be explicitly modeled and facilitates an analysis of the local dependence structure. Furthermore, the choice of the link function in Meier (2020) and Wang et al. (2022) might provide a poor approximation to the underlying distribution. In contrast, a result from Chernozhukov et al. (2018), guarantees that the joint distribution is locally Gaussian and BDR is able to approximate it.

We employ DR to investigate settings in which the outcome of interest is the joint distribution of two outcome variables conditional on a set of covariates. Moreover, we describe how bivariate distribution regression (BDR) can be implemented, as well as the potentially interesting functionals that can be derived from BDR estimates. We separately analyze the roles of the marginal distributions and the dependence structure. As each is modeled separately in BDR, a high degree of flexibility is incorporated between the role of both observable and unobservable factors. Alternative approaches to modeling conditional joint distributions exist. These include the use of copulas (i.e., Klein et al., 2022) and non-parametric estimators (i.e., Bouzebda and Nemouchi, 2019). Due to its semi-parametric properties, BDR can be seen as a mixture between these approaches. In contrast to copula models, BDR allows for greater flexibility in incorporating the impact of the covariates and requires relatively weaker parametric assumptions. Non-parametric approaches are more likely to suffer from the curse of dimensionality. A detailed overview of modeling choices for multivariate distributions can be found in Meier (2020).

Bivariate or multivariate distribution regression has been employed elsewhere. For example, Fernández-Val et al. (2023) model the joint distribution of spouses' wages in the presence of selection rules, Fernández-Val et al. (2024b) employ it in implementing difference-in-difference procedures with bivariate outcomes, while Fernández-Val et al. (2022) employ it to describe the joint distribution of spouses' annual earnings. The first two papers provide a theoretical contribution related to the literature on which they are focused and the third is purely a descriptive analysis. This paper provides important original theoretical contributions related to inference, the validity of the bootstrap employed, and the decompositions of the joint distribution and transition matrices.

Our empirical illustration is related to intergenerational mobility. This is a research area with potentially important policy implications on a range of topics such as tax policy and inequality reduction. It has received significant attention from policymakers and has generated a substantial academic literature. In a seminal paper, Chetty et al. (2014) showed that schools, neighborhoods, and family stability are influential determinants of mobility. Others have shown that race (Chetty et al., 2020), immigration status (Abramitzky et al., 2021), human capital (Adermon et al., 2021), or wealth (Adermon et al., 2018) are correlated with various mobility measures (see Mogstad and Torsvik (2023) for a recent overview). This suggests that one should include many covariates when modeling intergenerational mobility noting that the appropriate manner to do so is not straightforward given that the relationship between the covariates and earnings appears to vary over the earnings' distributions. Previous studies typically present estimates of relative mobility across subsamples of the data or employ a linear regression approach in which the child's outcomes (rank of income, human capital, or wealth) are regressed on the parents' analog and (sometimes) covariates. By employing BDR to model the joint distribution of fathers' and children's income we provide an additional methodological approach for evaluating intergenerational mobility. First, we allow for the covariates to have varying effects at different points of the earnings' distributions. Second, we can allow for different covariates to have an impact on each of the marginal distributions and the dependence structure. Third, as the joint distribution captures the entire dependence structure between the respective income measures, standard mobility measures, such as the rank-rank relationship and the conditional expected rank, can be directly derived from the BDR estimates. Fourth, we can explore how these mobility measures are influenced by the dependence in the unobservables and how this dependence may vary across the income distribution. Finally, BDR estimates can be employed to produce counterfactual bivariate distributions corresponding to scenarios capturing alternative values of the local correlation.

The remainder of this paper is organized as follows. Section 2 formally presents our methodological contribution. In Section 3 we present a number of functionals which are potentially of interest in empirical work. Section 4 outlines the associated estimation procedure. Section 5 presents the asymptotic theory. Section 6 contains our empirical application. Section 7 concludes.

2. Bivariate Distribution Regression

Let $F_{Y,W|X}$ be the joint distribution of (Y, W) conditional on X. We consider the bivariate distribution regression (BDR) model:

$$F_{Y,W|X}(y, w \mid x) = \Phi_2(x'\mu_y, x'\nu_w; g(x'\delta_{yw})), \tag{2.1}$$

where $\Phi_2(\cdot, \cdot; \rho)$ is the distribution of the standard bivariate normal with parameter ρ, and $u \mapsto g(u)$ is a known link function with range $[-1, 1]$ such as the Fisher transformation $g(u) = \tanh(u)$. In (2.1) the marginal distributions of Y and W conditional on X follow distribution regression models:

$$F_{Y|X}(y \mid x) = \Phi(x'\mu_y), \quad F_{W|X}(w \mid x) = \Phi(x'\nu_w),$$

where Φ is the distribution of the standard normal. The function $(y, w, x) \mapsto g(x'\delta_{yw})$ measures the local dependence or sorting between Y and W at $(Y, W, X) = (y, w, x)$. This sorting can vary with respect to observed covariates X, and along the distribution as indexed by (y, w). The simplest case is $g(x'\delta_{yw}) = g(\delta_{yw})$, where the sorting only varies with respect to unobservables.

The BDR model can be motivated by the local Gaussian representation (LGR) of Chernozhukov et al. (2018), which establishes that for any conditional joint distribution,

$$F_{Y,W|X}(y, w \mid x) = \Phi_2(\mu(y \mid x), \nu(w \mid x); \rho(y, w \mid x)), \tag{2.2}$$

for some functions μ, ν and ρ. The LGR is the right-hand-side of the previous equation and is unique, that is there is a one-to-one mapping between a conditional joint distribution and its LGR. In the LGR the marginal distributions of Y and W conditional on X are represented by nonparametric distribution regression models, that is

$$F_{Y|X}(y \mid x) = \Phi(\mu(y \mid x)), \quad F_{W|X}(w \mid x) = \Phi(\nu(w \mid x)).$$

The BDR model can be seen as a semiparametric specification for the LGR where the nonparametric functions μ, ν and ρ are replaced by (generalized) linear indexes with function-valued parameters. In particular, the parameters $y \mapsto \mu_y$ and $w \mapsto \nu_w$ measure the effect of the covariates on the marginal distributions of Y and W, and $(y, w) \mapsto \rho_{yw}$ measures the effect of the covariance on the local dependence (copula) between Y and W.

Let $\bar{\mathcal{Y}}$ and $\bar{\mathcal{W}}$ denote strict subsets of $\mathcal{Y}$ and $\mathcal{W}$, the supports of Y and W, respectively. We impose the following restrictions on the coefficients at the tails to facilitate estimation and inference,

$$\mu_{y,1} = \mu_{\bar{y}_y,1} + (y - \bar{y}_y)\alpha_{\bar{y}_y}, \quad \mu_{y,-1} = \mu_{\bar{y}_y,-1}, \quad y \in \mathcal{Y} \setminus \bar{\mathcal{Y}},$$

$$\nu_{w,1} = \nu_{\bar{w}_w,1} + (w - \bar{w}_w)\alpha_{\bar{w}_w}, \quad \nu_{w,-1} = \nu_{\bar{w}_w,-1}, \quad w \in \mathcal{W} \setminus \bar{\mathcal{W}},$$

and

$$\delta_{y,w} = \delta_{\bar{y}_y \bar{w}_w}, \quad y \in \mathcal{Y} \setminus \bar{\mathcal{Y}}, \quad w \in \mathcal{W} \setminus \bar{\mathcal{W}},$$

where $\bar{y}_y := \arg\min_{y' \in \bar{\mathcal{Y}}} |y - y'|$, $\alpha_{\bar{y}_y} > 0$, $\bar{w}_w := \arg\min_{w' \in \bar{\mathcal{W}}} |w - w'|$, $\alpha_{\bar{w}_w} > 0$, and the subscript 1 and -1 denote the first element of a vector and its complement. For example, $\mu_{y,1}$ is the first element of μ_y and $\mu_{y,-1}$ is a vector with the rest of the elements. Here, we follow Chernozhukov et al. (2025) and postulate that the random variables Y and W behave in the tails like random variables with distribution Φ, after subtracting the location shifts $x'\mu_{\bar{y}}$ and $x'\nu_{\bar{w}}$, and dividing by the scales $\alpha_{\bar{y}_y}$ and $\alpha_{\bar{w}_w}$, which are different at the upper and lower tails. For the local dependence parameter δ_{yw}, we postulate that it is constant at the tails. We could allow for additional tail parameters for the intercept of δ_{yw}, but we set them to zero for simplicity.[1]

3. Functionals of Interest

Rather than examine statistics which can be obtained in the univariate setting, we exploit the rich information contained in the conditional bivariate distribution and focus on functionals which highlight the dependence structure in the data. Accordingly, the local dependence parameter δ_{yw} features in each of the objects that follow.

3.1. Joint Distribution and Decomposition. The joint distribution of Y and W can be represented in terms of the BDR model as

$$F_{Y,W}(y, w) = \int \Phi_2(x'\mu_y, x'\nu_w; g(x'\delta_{yw}))\mathrm{d}F_X(x),$$

where F_X is the distribution of X. Using this representation, we can decompose the difference between the joint distribution of two groups, say $D = 1$ and $D = 0$. To do so, it is convenient to introduce counterfactual distributions that combine BDR parameters and distributions of X from different groups. Let

$$F_{Y,W}^{(j,k,l,m)}(y, w) := \int \Phi_2(x'\mu_y^j, x'\nu_w^k; g(x'\delta_{yw}^l))\mathrm{d}F_X^m(x), \tag{3.1}$$

where μ_y^i is the parameter μ_y in group i and the rest of the terms are defined analogously. In this notation the actual distribution in group d is $F_{Y,W}^{(d,d,d,d)}$.

The difference in the actual distributions between groups 1 and 0 can then be decomposed as

$$\underbrace{F_{Y,W}^{(1,1,1,1)}(y, w) - F_{Y,W}^{(0,0,0,0)}(y, w)}_{\text{Total}} = \underbrace{\left[F_{Y,W}^{(1,1,1,1)}(y, w) - F_{Y,W}^{(1,1,1,0)}(y, w)\right]}_{\text{Composition}}$$
$$+ \underbrace{\left[F_{Y,W}^{(1,1,1,0)}(y, w) - F_{Y,W}^{(1,1,0,0)}(y, w)\right]}_{\text{Sorting}} + \underbrace{\left[F_{Y,W}^{(1,1,0,0)}(y, w) - F_{Y,W}^{(0,0,0,0)}(y, w)\right]}_{\text{Marginals}}, \tag{3.2}$$

where the first term in brackets is a composition effect, the second term is a dependence or sorting effect, and the third term is the effect of the difference in the marginal distributions.

[1]Note that, unlike the intercepts of μ_y and ν_w, the intercept of δ_{yw} does not need to satisfy any monotonicity restriction at the tails.

The last term can be further decomposed into

$$\underbrace{\left[F_{Y,W}^{(1,1,0,0)}(y,w) - F_{Y,W}^{(1,0,0,0)}(y,w)\right]}_{\text{Marginal of } W} + \underbrace{\left[F_{Y,W}^{(1,0,0,0)}(y,w) - F_{Y,W}^{(0,0,0,0)}(y,w)\right]}_{\text{Marginal of } Y}, \tag{3.3}$$

where the first term is the marginal effect related to W and the last term is the marginal effect related to W.

3.2. Transition Matrices and Decomposition. We can also construct transition matrices and decompose differences across those for different groups into composition, sorting and marginal distribution components using counterfactual distributions. Let $\{y_j : 0 \leqslant j \leqslant J\}$ and $\{w_k : 0 \leqslant k \leqslant K\}$ be grids of values covering the supports of Y and W, where we set $y_0 = w_0 = -\infty$ and $y_K = w_K = +\infty$. Then, the (j,k) element of the counterfactual transition matrix $T^{(k,d,r,s)}$ is

$$T_{jk}^{(k,d,r,s)} := F_{Y,W}^{(k,d,r,s)}(y_j, w_k) - F_{Y,W}^{(k,d,r,s)}(y_{j-1}, w_k) - F_{Y,W}^{(k,d,r,s)}(y_j, w_{k-1}) + F_{Y,W}^{(k,d,r,s)}(y_{j-1}, w_{k-1}). \tag{3.4}$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$. These matrices provide a parsimonious representation of the joint distribution of Y and W.

The difference in the actual transition matrices between groups 1 and 0 can then be decomposed as

$$\begin{aligned}\underbrace{T^{(1,1,1,1)} - T^{(0,0,0,0)}}_{\text{Total}} = \underbrace{\left[T^{(1,1,1,1)} - T^{(1,1,1,0)}\right]}_{\text{Composition}} + \underbrace{\left[T^{(1,1,1,0)} - T^{(1,1,0,0)}\right]}_{\text{Sorting}} \\ + \underbrace{\left[T^{(1,1,0,0)} - T^{(1,0,0,0)}\right]}_{\text{Marginal of } W} + \underbrace{\left[T^{(1,0,0,0)} - T^{(0,0,0,0)}\right]}_{\text{Marginal of } Y},\end{aligned}$$

We provide examples of this decomposition in Section 6.

4. Estimation

Rather than employing probit, as is done in univariate DR, BDR estimates a sequence of bivariate probits. Let $I^y := 1(Y \leqslant y)$, $\bar{I}^y := 1 - I^y$, $J^w := (W \leqslant w)$ and $\bar{J}^w := 1 - J^w$, for $(y,w) \in \mathcal{Y}_n \times \mathcal{W}_n$, where $\mathcal{Y}_n$ and $\mathcal{W}_n$ are grids of points in the supports of Y and W, respectively.

Under the BDR model, the joint probability function of I^y and J^w conditional on $X = x$ is

$$\begin{aligned} f_{I^y,J^w|X}(i,j \mid x,\mu_y,\nu_w,\delta_{y,w}) = {} & \Phi_2(x'\mu_y, x'\nu_w; \rho(x'\delta_{yw})^{ij} \\ & \times \Phi_2(x'\mu_y, -x'\nu_w; -\rho(x'\delta_{yw})^{i(1-j)} \\ & \times \Phi_2(-x'\mu_y, x'\nu_w; -\rho(x'\delta_{yw})^{(1-i)j} \\ & \times \Phi_2(-x'\mu_y, -x'\nu_w; \rho(x'\delta_{yw})^{(1-i)(1-j)} \\ & i,j \in \{0,1\}. \end{aligned} \tag{4.1}$$

Given a random sample of size n, $\{(Y_i, W_i, X_i)\}_{=1}^n$, the average conditional log-likelihood at (y, w) is

$$\ell^{yw}(\mu,\nu,\delta) = \frac{1}{n}\sum_{i=1}^n \log f_{I^c,J^f|Z}(I_i^c, J_i^f \mid X_i, \mu, \nu, \delta). \tag{4.2}$$

The conditional maximum likelihood estimator (CMLE) of $(\mu_y, \nu_w, \delta_{yw})$ is

$$(\widehat{\mu}_y, \widehat{\nu}_w, \widehat{\delta}_{yw}) \in \arg\max_{\mu,\nu,\delta} \ell^{y,w}(\mu,\nu,\delta). \tag{4.3}$$

Several remarks should be noted regarding computation. First, when $\rho(x'\delta_{yw}) = \rho(\delta_{yw})$, then the CMLE is the bivariate probit estimator of (I^y, J^w) on X. Second, the CMLE can be computed in two steps. In the first step, we obtain the estimators of μ_y and ν_w by probit distribution regression of I^y on X and J^w on X, respectively. In the second step, we plug in the estimators of μ_y and ν_w from the first step in the average log-likelihood and maximize with respect to δ, that is

$$\widehat{\delta}_{yw} \in \arg\max_{\delta} \ell^{y,w}(\widehat{\mu}_y, \widehat{\nu}_w, \delta), \tag{4.4}$$

where $\widehat{\mu}_y$ and $\widehat{\nu}_w$ are the distribution regression estimators of μ_y and ν_w. Finally, to trace the joint and marginal distributions of Y and W, we need to obtain the CMLE for a grid of values of (y, w) in the support of (Y, W). For example, we can use the tensor product of two grids for Y and W, including sampling percentiles from the 1 or 2% to the 99 or 98%.

Let $\mathbb{E}_n$ denote the empirical expectation, that is $\mathbb{E}_n := n^{-1}\sum_{i=1}$.

Algorithm 1 (BDR Estimator)**.** Let $d_x := \dim X$, $\mathbb{R}_n$ denote the set containing the observed values of R and $\bar{\mathbb{R}}_n = \mathbb{R}_n \cap \bar{\mathbb{R}}$, for $R \in \{Y, W\}$.

(1) Estimate μ_y at $y \in \bar{\mathcal{Y}}_n$ and ν_w at $w \in \bar{\mathcal{W}}_n$ by DR, that is,

$$\widehat{\mu}_y \in \arg\max_{\mu \in \mathbb{R}^{d_x}} \ell^y(\mu) = \mathbb{E}_n[\ell_i^y(\mu)], \quad \ell_i^y(\mu) := I_i^y \log\Phi(X_i'\mu) + \bar{I}_i^y \log\Phi(-X_i'\mu),$$

and

$$\widehat{\nu}_w \in \arg\max_{\nu \in \mathbb{R}^{d_x}} \ell^w(\mu) = \mathbb{E}_n[\ell^w_i(\nu)], \quad \ell^w_i(\mu) := J^w_i \log \Phi(X_i'\nu) + \bar{J}^w_i \log \Phi(-X_i'\nu).$$

(2) Estimate μ_y at $y \in \mathcal{Y}_n \setminus \bar{\mathcal{Y}}_n$ and ν_w at $w \in \mathcal{W}_n \setminus \bar{\mathcal{W}}_n$ by restricted DR, that is,

$$\widehat{\mu}_y = (y - \bar{y}_y)\widehat{\alpha}_{\bar{y}_y} + x'\widehat{\mu}_{\bar{y}_y} \text{ and } \widehat{\nu}_w = (w - \bar{w}_w)\widehat{\alpha}_{\bar{w}_w} + x'\widehat{\nu}_{\bar{w}_w},$$

where $\bar{y}_y := \arg\min_{y' \in \bar{\mathcal{Y}}_n} |y - y'|$, $\bar{w}_w := \arg\min_{w' \in \bar{\mathcal{W}}_n} |w - w'|$,

$$\begin{aligned} \widehat{\alpha}_{\bar{y}_y} &\in \arg\max_{a \in \mathbb{R}} \mathbb{E}_n[\ell^{y_0}_i(a, \widehat{\mu}_{\bar{y}_y})], \\ \ell^{y_0}_i(a, \mu) &:= \left[I^{y_0}_i \log \Phi((y_0 - \bar{y}_y)a + X_i'\mu) + \bar{I}^{y_0}_i \log\{\Phi(-(y_0 - \bar{y}_y)a - X_i'\mu)\right], \\ \widehat{\alpha}_{\bar{w}_w} &\in \arg\max_{a \in \mathbb{R}} \mathbb{E}_n[\ell^{w_0}_i(a, \widehat{\nu}_{\bar{w}_w})], \\ \ell^{w_0}_i(a, \nu) &:= \left[J^{w_0}_i \log \Phi((w_0 - \bar{w}_w)a + X_i'\nu) + \bar{J}^{w_0}_i \log\{\Phi(-(w_0 - \bar{w}_w)a - X_i'\nu)\right], \end{aligned}$$

and, for $r \in \{y, w\}$ and $\mathcal{R} \in \{\mathcal{Y}, \mathcal{W}\}$, $r_0 \in \mathbb{R}_n \setminus \bar{\mathbb{R}}_n$ is such that (i) there are at least m observations between $\bar{r}$ and r_0, and greater than r_0 if $r_0 > \bar{r}_r$ (upper tail) or less than r_0 if $r_0 < \bar{r}_r$ (lower tail), and (ii) $\widehat{\alpha}_{\bar{r}_r} > 0$.

(3) Estimate δ_{yw} at $(y, w) \in \bar{\mathcal{Y}}_n \times \bar{\mathcal{W}}_n$ by restricted BDR, that is,

$$\widehat{\delta}_{yw} \in \arg\max_{\delta \in \mathbb{R}^{d_x}} \ell^{yw}(\widehat{\mu}_y, \widehat{\nu}_w, \delta) = \mathbb{E}_n[\ell^{yw}_i(\widehat{\mu}_y, \widehat{\nu}_w, \delta)],$$

where

$$\begin{aligned} \ell^{yw}_i(\mu, \nu, \delta) &= I^y_i J^w_i \log \Phi_2(X_i'\mu, X_i'\nu; g(X_i'\delta) + I^y_i \bar{J}^w_i \log \Phi_2(X_i'\mu, -X_i'\nu; -g(X_i'\delta) \\ &+ \bar{I}^y_i J^w_i \log \Phi_2(-X_i'\mu, X_i'\nu; -g(X_i'\delta)) + \bar{I}^y_i \bar{J}^w_i \log \Phi_2(-X_i'\mu, -X_i'\nu; g(X_i'\delta)). \end{aligned}$$

(4) Estimate δ_{yw} at $(y, w) \in (\mathcal{Y}_n \setminus \bar{\mathcal{Y}}_n) \times (\mathcal{W}_n \setminus \bar{\mathcal{W}}_n)$ by imposing the tail restrictions, that is,

$$\widehat{\delta}_{yw} = \widehat{\delta}_{\bar{y}_y \bar{w}_w}.$$

The counterfactual terms of the decomposition of the joint distribution can be estimated via a plug-in rule. For example, an estimator of the second integral on the right-hand side of (3.2) can be constructed as follows:

$$\frac{1}{n_0} \sum_{i=1}^{n} (1 - D_i) \Phi_2(X_i'\widehat{\mu}^1_y, X_i'\widehat{\nu}^1_w; g(X_i'\widehat{\delta}^1_{yw})),$$

where $\widehat{\mu}^1_y, \widehat{\nu}^1_w, \widehat{\delta}^1_{yw}$ are estimators of $\mu^1_y, \nu^1_w, \delta^1_{yw}$ in the subsample with $D = 1$ and $n_0 := \sum_{i=1}^n (1 - D_i)/n$. [2]

[2]Alternatively, a doubly-robust estimator can be formed as

$$\frac{1}{n_0} \sum_{i=1}^{n} \frac{(1 - D_i)}{1 - \widehat{p}(X_i)} 1\{Y_i \leqslant y, W_i \leqslant w\} - \frac{1}{n_0} \sum_{i=1}^{n} \frac{(D_i - \widehat{p}(X_i))}{1 - \widehat{p}(X_i)} \Phi_2(X_i'\widehat{\mu}^1_y, X_i'\widehat{\nu}^1_w; g(X_i'\widehat{\delta}^1_{yw})),$$

4.1. Inference. We use the exchangeable bootstrap for inference. The following algorithm describes how to obtain counterfactual draws of the BDR estimator.

Algorithm 2 (Exchangeable Bootstrap Draw of BDR Estimator)**.**

(1) Draw a realization of the weights $(\omega_{n1}, \ldots, \omega_{nn})$ from a distribution that satisfies Assumption 5.2 in Section 5. Normalize the weights to add-up to one.

(2) Obtain a bootstrap draw of $\widehat{\mu}_y$ at $y \in \bar{\mathcal{Y}}_n$ and $\widehat{\nu}_w$ at $w \in \bar{\mathcal{W}}_n$ by weighted DR, that is,

$$\widehat{\mu}_y^* \in \arg\max_{\mu \in \mathbb{R}^{d_x}} \ell_*^y(\mu) = \mathbb{E}_n[\omega_{ni} \ell_i^y(\mu)], \quad \widehat{\nu}_w^* \in \arg\max_{\nu \in \mathbb{R}^{d_x}} \ell_*^w(\mu) = \mathbb{E}_n[\omega_{ni} \ell_i^w(\nu)].$$

(3) Obtain bootstrap draw of $\widehat{\mu}_y$ at $y \in \mathcal{Y}_n \setminus \bar{\mathcal{Y}}_n$ and $\widehat{\nu}_w$ at $w \in \mathcal{W}_n \setminus \bar{\mathcal{W}}_n$ by weighted restricted DR, that is,

$$\widehat{\mu}_y^* = (y - \bar{y}_y)\widehat{\alpha}_{\bar{y}_y}^* + x'\widehat{\mu}_{\bar{y}_y}^* \text{ and } \widehat{\nu}_w^* = (w - \bar{w}_w)\widehat{\alpha}_{\bar{w}_w}^* + x'\widehat{\nu}_{\bar{w}_w}^*,$$

where

$$\widehat{\alpha}_{\bar{y}_y}^* \in \arg\max_{a \in \mathbb{R}} \mathbb{E}_n[\omega_{ni} \ell_i^{y_0}(a, \widehat{\mu}_{\bar{y}_y}^*)], \quad \widehat{\alpha}_{\bar{w}_w}^* \in \arg\max_{a \in \mathbb{R}} \mathbb{E}_n[\omega_{ni} \ell_i^{w_0}(a, \widehat{\nu}_{\bar{w}_w}^*)],$$

and, for $r \in \{y, w\}$ and $\mathcal{R} \in \{\mathcal{Y}, \mathcal{W}\}$, $r_0 \in \mathbb{R}_n \setminus \bar{\mathbb{R}}_n$ is the same as in Algorithm 1(2).

(4) Obtain bootstrap draw of $\widehat{\delta}_{yw}$ at $(y, w) \in \bar{\mathcal{Y}}_n \times \bar{\mathcal{W}}_n$ by weighted restricted BDR, that is,

$$\widehat{\delta}_{yw}^* \in \arg\max_{\delta \in \mathbb{R}^{d_x}} \ell_*^{yw}(\widehat{\mu}_y, \widehat{\nu}_w, \delta) = \mathbb{E}_n[\omega_{ni} \ell_i^{yw}(\widehat{\mu}_y, \widehat{\nu}_w, \delta)].$$

(5) Obtain bootstrap draw of $\widehat{\delta}_{yw}$ at $(y, w) \in (\mathcal{Y}_n \setminus \bar{\mathcal{Y}}_n) \times (\mathcal{W}_n \setminus \bar{\mathcal{W}}_n)$ by imposing the tail restrictions, that is,

$$\widehat{\delta}_{yw}^* = \widehat{\delta}_{\bar{y}_y \bar{w}_w}^*.$$

Exchangeable bootstrap draws of the estimators of the functionals of interest can be obtained similarly.

5. Asymptotic Theory

To simplify the expressions of some of the assumptions and limit processes, it is convenient to introduce the following notation for the tails. Let $\bar{r}_r := \arg\min_{r' \in \bar{\mathcal{R}}} |r - r'|$ for $r \in \{y, w\}$ and $\bar{\mathcal{R}} \in \{\bar{\mathcal{Y}}, \bar{\mathcal{W}}\}$. Note that $\bar{r}_r = r$ if $r \in \bar{\mathcal{R}}$, $\bar{r}_r = \bar{r}$ if $r \geq \bar{r}$ and $\bar{r}_r = \underline{r}$ if $r \leqslant \underline{r}$, where $\bar{r} := \sup(\bar{\mathcal{R}})$ and $\underline{r} := \inf(\bar{\mathcal{R}})$. We use a superscript d to denote variables, parameters and samples from the group $d \in \{0, 1\}$. We also use the following notation for partial derivatives: $\partial_x f(x) := \partial f(x)/\partial x$ and $\partial_{xx} f(x) := \partial^2 f(x)/(\partial x \partial x')$.

---

where $\widehat{p}(x)$ is an estimator of the propensity score $p(x) = \mathrm{P}(D = 1 \mid X = x)$.

Assumption 5.1 (Model and Sampling)**.** *For $d \in \{0,1\}$: (1) Random sampling: $\{Z_i^d := (Y_i^d, W_i^d, X_i^d)\}_{i=1}^{n_d}$ is a sequence of independent and identically distributed copies of $Z^d := (Y^d, W^d, X^d)$, which is independent across d, and $n_d/n \to p_d > 0$ for $n := n_0 + n_1$.*[3] *(2) Model: the joint distribution of (Y^d, W^d) conditional on X^d follows the BDR model (2.1) with the tail restrictions. (3) The support of X^d, $\mathcal{X}$, is a compact set. (4) The supports of Y and W, $\mathcal{Y}$ and $\mathcal{W}$ are either (i) finite sets or (ii) open intervals. In the second case, the conditional density functions $f_{Y^d|X^d}(y \mid x)$ and $f_{W^d|X^d}(w \mid x)$ exist, are uniformly bounded above, and are uniformly continuous in (y,x) on $\mathcal{Y} \times \mathcal{X}$ and in (w,x) on $\mathcal{W} \times \mathcal{X}$, respectively. (5) Identification and non-degeneracy: the equation $\mathrm{E}[\partial_\delta \ell_i^{yw}(\mu_y^d, \nu_w^d, \tilde{\delta}_{yw})] = 0$ possesses a unique solution at $\tilde{\delta}_{yw} = \delta_{yw}^d$ that lies in the interior of a compact set $\mathcal{D} \subset \mathbb{R}^{d_\delta}$ for all $(y,w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}} := \bar{\mathcal{Y}} \times \bar{\mathcal{W}}$; the minimum eigenvalues of the matrices $\mathrm{E}[\partial_{\mu\mu}\ell_i^y(\mu_y^d)]$, $\mathrm{E}[\partial_{\alpha\alpha}\ell_i^{y_0}(\alpha_{\bar{y}_y}^d, \mu_{\bar{y}_y})]$, $\mathrm{E}[\partial_{\nu\nu}\ell_i^w(\nu_w^d)]$, $\mathrm{E}[\partial_{\alpha\alpha}\ell_i^{w_0}(\alpha_{\bar{w}_w}^d, \mu_{\bar{w}_w})]$ and $\mathrm{E}[\partial_{\delta\delta}\ell_i^{yw}(\mu_y^d, \nu_w^d, \delta_{yw}^d)]$ are bounded away from zero uniformly over $y \in \bar{\mathcal{Y}}$, $y \in \{\underline{y}, \bar{y}\}$, $w \in \bar{\mathcal{W}}$, $w \in \{\underline{w}, \bar{w}\}$, and $(y,w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}}$, respectively; and $\mathrm{E}\|X\|^2 < \infty$.*

Assumption 5.1 implicitly imposes common support for the variables Y^d, W^d and X^d across d. This condition guarantees that all the counterfactual distributions that we consider are well-defined.

Assumption 5.2 (Exchangeable Bootstrap)**.** *For each n_d and $d \in \{0,1\}$, $(\omega_{n_d 1}^d, ..., \omega_{n_d n_d}^d)$ is an exchangeable,*[4] *nonnegative random vector, which is independent of the data and across d, such that for some $\epsilon > 0$*

$$\sup_n \mathrm{E}[(\omega_{n_d 1}^d)^{2+\epsilon}] < \infty, \;\; n_d^{-1}\sum_{i=1}^{n_d}\left(\omega_{n_d i}^d - \bar{\omega}_{n_d}^d\right)^2 \to_{\mathrm{P}} 1, \;\; \bar{\omega}_{n_d}^d \to_{\mathrm{P}} 1, \tag{5.1}$$

where $\bar{\omega}_{n_d}^d = n_d^{-1}\sum_{i=1}^{n_d}\omega_{n_d i}^d$.

Let $\widehat{\theta}_{yw}^d := (\widehat{\mu}_y^{d\prime}, \widehat{\nu}_w^{d\prime}, \widehat{\delta}_{yw}^{d\prime})'$ and $\widehat{\theta}_{yw}^{d*} := (\widehat{\mu}_y^{d*\prime}, \widehat{\nu}_w^{d*\prime}, \widehat{\delta}_{yw}^{d*\prime})'$ be the estimator and corresponding bootstrap draw of the parameter $\theta_{yw}^d := (\mu_y^{d\prime}, \nu_w^{d\prime}, \delta_{yw}^{d\prime})'$, for $y \in \bar{\mathcal{Y}}$ and $w \in \bar{\mathcal{W}}$. Similarly, let $\widehat{\xi}_y^d := (\widehat{\alpha}_y^d, \widehat{\mu}_y^{d\prime})'$ and $\widehat{\xi}_w^d := (\widehat{\alpha}_w^d, \widehat{\nu}_w^{d\prime})'$ be the estimators of the tail parameters $\xi_y^d := (\alpha_y^d, \mu_y^{d\prime})'$ and $\xi_w^d := (\alpha_w^d, \nu_w^{d\prime})'$, for $y \in \{\underline{y}, \bar{y}\}$ and $w \in \{\underline{w}, \bar{w}\}$, and $\widehat{\xi}_y^{d*} := (\widehat{\alpha}_y^{d*}, \widehat{\mu}_y^{d*\prime})'$ and $\widehat{\xi}_w^{d*} := (\widehat{\alpha}_w^{d*}, \widehat{\nu}_w^{d*\prime})'$ be the corresponding bootstrap draws of the estimators. Let $Z_n \rightsquigarrow Z$ in $\mathbb{D}$ denote weak convergence of a stochastic process Z_n to a random element Z in a normed space $\mathbb{D}$, as defined in van der Vaart and Wellner (1996). For example, $\mathbb{D} = \ell^\infty(\bar{\mathcal{Y}}\bar{\mathcal{W}})$, the set of measurable and bounded functions on $\bar{\mathcal{Y}}\bar{\mathcal{W}}$.

[3] Some forms of dependence between the samples across d can be accommodated following the analysis in Chernozhukov et al. (2013). We assume independence for simplicity and because it holds in our application.

[4] A sequence of random variables $X_1, X_2, ..., X_n$ is exchangeable if for any finite permutation σ of the indices $1, 2, ..., n$ the joint distribution of the permuted sequence $X_{\sigma(1)}, X_{\sigma(2)}, ..., X_{\sigma(n)}$ is the same as the joint distribution of the original sequence.

In order to state the results about bootstrap validity formally, we follow the notation and definitions in van der Vaart and Wellner (1996). Let D_n denote the data vector and M_n be the vector of random variables used to generate bootstrap draws given D_n. Consider the random element $\mathbb{Z}_n^* = \mathbb{Z}_n(D_n, M_n)$ in a normed space $\mathbb{E}$. We say that the bootstrap law of $\mathbb{Z}_n^*$ consistently estimates the law of some tight random element $\mathbb{Z}$ and write $\mathbb{Z}_n^* \rightsquigarrow_{\mathrm{P}} \mathbb{Z}$ in $\mathbb{D}$ if:

$$\sup_{h \in \mathrm{BL}_1(\mathbb{D})} |\mathrm{E}_{M_n} h(\mathbb{Z}_n^*) - \mathrm{E}h(\mathbb{Z})| \to_{\mathrm{P}} 0, \tag{5.2}$$

where $\mathrm{BL}_1(\mathbb{D})$ denotes the space of functions with Lipschitz norm at most 1 and E_{M_n} denotes the conditional expectation with respect to M_n given the data D_n; and $\to_{\mathrm{P}}$ denotes convergence in (outer) probability.

Lemma 5.1 (FCLT and Bootstrap FCLT for $(\widehat{\theta}_{yw}^d, \widehat{\xi}_{\bar{y}}^d, \widehat{\xi}_{\underline{y}}^d, \widehat{\xi}_{\bar{w}}^d, \widehat{\xi}_{\underline{w}}^d)$). *Assume that Assumption 5.1 holds. Then, for $d \in \{0,1\}$, (i)*

$$\sqrt{n}(\widehat{\theta}_{yw}^d - \theta_{yw}^d) \rightsquigarrow Z_{yw}^{d\theta} := \sqrt{p_d}\,\mathbb{G}[\psi_i(\theta_{yw}^d)] \text{ in } \ell^\infty(\bar{\mathcal{Y}}\bar{\mathcal{W}})^{d_\theta},$$

where $\mathbb{G}$ is a Brownian bridge and $\psi_i(\theta_{yw}^d) := H^{yw}(\theta_{yw}^d)^{-1} S_i^{yw}(\theta_{yw}^d)$ with

$$S_i^{yw}(\theta_{yw}^d) = \begin{bmatrix} \partial_\mu \ell_i^y(\mu_y^d) \\ \partial_\nu \ell_i^w(\nu_w^d) \\ \partial_\delta \ell_i^{yw}(\mu_y^d, \nu_w^d, \delta_{yw}^d) \end{bmatrix}, \tag{5.3}$$

and

$$H^{yw}(\theta_{yw}^d) = \mathrm{E}\begin{bmatrix} \partial_{\mu\mu} \ell_i^y(\mu_y^d) & 0 & 0 \\ 0 & \partial_{\nu\nu} \ell_i^w(\nu_w^d) & 0 \\ \partial_{\delta\mu} \ell_i^{yw}(\mu_y^d, \nu_w^d, \delta_{yw}^d) & \partial_{\delta\nu} \ell_i^{yw}(\mu_y^d, \nu_w^d, \delta_{yw}^d) & \partial_{\delta\delta} \ell_i^{yw}(\mu_y^d, \nu_w^d, \delta_{yw}^d) \end{bmatrix}; \tag{5.4}$$

and (ii) $\sqrt{n}(\widehat{\xi}_r^d - \xi_r^d) \rightsquigarrow Z^{d\xi_r}$ in $\mathbb{R}^{d_{\xi_r}}$, for $r \in \{\bar{r}, \underline{r}\}$ and $r \in \{y, w\}$, where

$$Z^{d\xi_r} := \sqrt{p_d}\,\mathrm{E}\begin{bmatrix} \partial_{\alpha\alpha} \ell_i^{r_0}(\alpha_{\bar{r}_r}^d, \mu_{\bar{r}_r}^d) & \partial_{\alpha\mu} \ell_i^{r_0}(\alpha_{\bar{r}_r}^d, \mu_{\bar{r}_r}^d) \\ 0 & \partial_{\mu\mu} \ell_i^y(\mu_{\bar{r}_r}^d) \end{bmatrix}^{-1} \mathbb{G}\begin{bmatrix} \partial_\alpha \ell_i^{r_0}(\alpha_{\bar{r}_r}^d, \mu_{\bar{r}_r}^d) \\ \partial_\mu \ell_i^r(\mu_{\bar{r}_r}^d) \end{bmatrix},$$

has a zero-mean normal distribution. Moreover, (i) and (ii) hold jointly, and $(Z_{yw}^{0\theta}, Z^{0\xi_{\bar{r}}}, Z^{0\xi_{\underline{r}}})$ and $(Z_{yw}^{1\theta}, Z^{1\xi_{\bar{r}}}, Z^{1\xi_{\underline{r}}})$ are independent. If in addition Assumption 5.2 holds, for $d \in \{0,1\}$, jointly (i^*)

$$\sqrt{n}(\widehat{\theta}_{yw}^{d*} - \widehat{\theta}_{yw}^d) \rightsquigarrow_{\mathrm{P}} Z_{yw}^{d\theta} \text{ in } \ell^\infty(\bar{\mathcal{Y}}\bar{\mathcal{W}})^{d_\theta},$$

and (ii^*) $\sqrt{n}(\widehat{\xi}_r^{d*} - \widehat{\xi}_r^d) \rightsquigarrow_{\mathrm{P}} Z^{d\xi_r}$ in $\mathbb{R}^{d_{\xi_r}}$, for $r \in \{\bar{r}, \underline{r}\}$ and $r \in \{y, w\}$.

The quantities of interest are functionals of the joint distribution of (Y, W) conditional on X and the marginal distribution of X. To state the results about these functionals it is convenient to introduce some notation. The following definitions apply to all $j, k, l, m \in \{0, 1\}$.

Let F_X^m, $\widehat{F}_X^m$ and $\widehat{F}_X^{m*}$ be the joint distribution of X^m, empirical counterpart and bootstrap draw; and $F_{ywx}^{(jkl)} := \Phi_2(x'\mu_y^j, x'\nu_w^k; g(x'\delta_{yw}^l))$, $\widehat{F}_{ywx}^{(j,k,l)} := \Phi_2(x'\widehat{\mu}_y^j, x'\widehat{\nu}_w^k; g(x'\widehat{\delta}_{yw}^l))$ and $\widehat{F}_{ywx}^{(jkl)*} := \Phi_2(x'\widehat{\mu}_y^{j*}, x'\widehat{\nu}_w^{k*}; g(x'\widehat{\delta}_{yw}^{l*}))$ be the counterfactual conditional distribution, its estimator and bootstrap draw. Consider the empirical processes and their bootstrap draws, $(y, w, x) \mapsto \widehat{Z}_{ywx}^{(jkl)} := \sqrt{n}(\widehat{F}_{ywx}^{(jkl)} - F_{ywx}^{(jkl)})$, $(y, w, x) \mapsto \widehat{Z}_{ywx}^{(jkl)*} := \sqrt{n}(\widehat{F}_{ywx}^{(jkl)*} - \widehat{F}_{ywx}^{(jkl)})$, $f \mapsto \widehat{G}_X^m(f) := \sqrt{n}\int f \mathrm{d}(\widehat{F}_X^m - F_X^m)$ and $f \mapsto \widehat{G}_X^{m*}(f) := \sqrt{n}\int f \mathrm{d}(\widehat{F}_X^{m*} - \widehat{F}_X^m)$, for $f \in \mathcal{F}$, where $\mathcal{F}$ is a class of measurable functions that (i) includes $F_{ywx}^{(jkl)}$, the indicators of all the rectangles in $\bar{\mathbb{R}}^{d_x+2}$, for $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, the extended real line, and (ii) is totally bounded under the metric:

$$\lambda(f; \tilde{f}) = \left[\int (f - \tilde{f})\mathrm{d}F_X\right]^{1/2}, \quad f, \tilde{f} \in \mathcal{F}.$$

For $d \in \{0, 1\}$, the selection matrix $\boldsymbol{S}_d^{(jkl)}$ picks up the elements of $Z_{ywx}^{d\theta}$ corresponding to the components used in the construction of the counterfactual conditional distribution. For example, $\boldsymbol{S}_0^{(010)} = \mathrm{diag}(\boldsymbol{I}_{d_\mu}, \boldsymbol{0}_{d_\nu}, \boldsymbol{I}_{d_\delta})$ and $\boldsymbol{S}_1^{(010)} = \mathrm{diag}(\boldsymbol{0}_{d_\mu}, \boldsymbol{I}_{d_\nu}, \boldsymbol{0}_{d_\delta})$, where $\boldsymbol{0}_p$ is a $p \times p$ matrix of zeros and $\boldsymbol{I}_p$ is the identity matrix of size p.

Theorem 5.1 (FCLT and Bootstrap CLT for $\widehat{F}_{ywx}^{(jkl)}$ and $\widehat{G}_X^m(f)$)**.** (*i*) *Assume that Assumption 5.1 holds. Then, jointly in $j, k, l, m \in \{0, 1\}$,*

$$(\widehat{Z}_{ywx}^{(jkl)}, \widehat{G}_X^m(f)) \rightsquigarrow (Z_{ywx}^{(jkl)}, G_X^m(f)) \text{ in } \ell^\infty(\mathcal{YWXF}),$$

where $G_X^m(f) := \sqrt{p_m}\mathbb{G}^m(f)$, with $\mathbb{G}^0$ and $\mathbb{G}^1$ independent, and

$$Z_{ywx}^{(jkm)} := \partial_\theta F_{\bar{y}_y\bar{w}_w x}^{(jkl)} \sum_{d\in\{0,1\}} \boldsymbol{S}_d^{(jkl)} Z_{\bar{y}_y\bar{w}_w}^{d\theta} + \partial_{\alpha_y} F_{\bar{y}_y\bar{w}_w x}^{(jkl)} Z^{j\alpha_{\bar{y}_y}} 1(y \notin \bar{\mathcal{Y}}) + \partial_{\alpha_w} F_{\bar{y}_y\bar{w}_w x}^{(jkl)} Z^{k\alpha_{\bar{w}_w}} 1(w \notin \bar{\mathcal{W}}),$$

with $\partial_\theta F_{\bar{y}_y\bar{w}_w x}^{(jkl)} := \partial_\theta \Phi_2(x'\mu_{\bar{y}_y}^j, x'\nu_{\bar{w}_w}^k; g(x'\delta_{\bar{y}_y\bar{w}_w}^l))$, $\partial_{\alpha_y} F_{\bar{y}_y wx}^{(jkl)} := \partial_{\alpha_{\bar{y}_y}} \Phi_2(\alpha_{\bar{y}_y}^j(y - \bar{y}_y) + x'\mu_{\bar{y}_y}^j, x'\nu_w^k; g(x'\delta_{yw}^l))$, $\partial_{\alpha_w} F_{y\bar{w}_w x}^{(jkl)} := \partial_{\alpha_{\bar{w}_w}} \Phi_2(x'\mu_y^j, \alpha_{\bar{w}_w}^l(w - \bar{w}_w) + x'\nu_{\bar{w}_w}^l; g(x'\delta_{yw}^l))$, $Z^{d\alpha_r} := e_1' Z^{d\xi_r}$ for $r \in \{y, w\}$ and $d \in \{0, 1\}$, e_1 is a unitary vector with a one in the first position, and $Z_{yw}^{d\theta}$, $Z^{d\xi_y}$ and $Z^{d\xi_w}$ are defined in Lemma 5.1. (ii) If in addition Assumption 5.2 holds, then, jointly in $j, k, l, m \in \{0, 1\}$,

$$(\widehat{Z}_{ywx}^{(jkl)*}, \widehat{G}_Z^{m*}(f)) \rightsquigarrow_{\mathrm{P}} (Z_{ywx}^{(jkl)}, G_Z^m(f)) \text{ in } \ell^\infty(\mathcal{YWXF}).$$

Let

$$\widehat{F}_{Y,W}^{(j,k,l,m)}(y, w) = \frac{1}{n_m}\sum_{i=1}^{n_m} \Phi_2(X_i^{m\prime}\widehat{\mu}_y^j, X_i^{m\prime}\widehat{\nu}_w^k; g(X_i^{m\prime}\widehat{\delta}_{yw}^l)),$$

be an estimator of the counterfactual joint distribution in (3.1) and

$$\widehat{F}^{*(j,k,l,m)}_{Y,W}(y,w) = \frac{1}{n_m}\sum_{i=1}^{n_m} \omega^m_{in_m}\Phi_2(X_i^{m\prime}\widehat{\mu}^{j*}_y, X_i^{m\prime}\widehat{\nu}^{k*}_w; g(X_i^{m\prime}\widehat{\delta}^{l*}_{yw})),$$

be the corresponding bootstrap draw. The following result follows from Theorem 5.1, Lemma D.1 of Chernozhukov et al. (2013), which establish the Hadamard differentiability of the counterfactual map, and the functional delta method.

Corollary 5.1 (FCLT and Bootstrap FCLT for $\widehat{F}^{(j,k,l,m)}_{Y,W}$). *Under the assumptions of Theorem 5.1, jointly for $i,j,k,l \in \{0,1\}$,*

$$\sqrt{n}(\widehat{F}^{(j,k,l,m)}_{Y,W}(y,w) - F^{(j,k,l,m)}_{Y,W}(y,w)) \rightsquigarrow Z^{(j,k,l,m)}_{yw} := \int Z^{(jkl)}_{ywx}\mathrm{d}F^m_X(x) + G^m_X(F^{(jkl)}_{ywx}) \text{ in } \ell^\infty(\mathcal{YW}),$$

and

$$\sqrt{n}(\widehat{F}^{*(j,k,l,m)}_{Y,W}(y,w) - \widehat{F}^{(j,k,l,m)}_{Y,W}(y,w)) \rightsquigarrow_{\mathrm{P}} Z^{(j,k,l,m)}_{yw} \text{ in } \ell^\infty(\mathcal{YW}).$$

6. Empirical Illustration

We now employ BDR to study intergenerational income mobility. BDR is a particularly useful and powerful tool for this area of research since it can model the joint distribution of incomes for different generations of family members. We focus on the joint distribution of a child's and their father's labor incomes. We begin by documenting the raw data. This provides some insight into the presence of intergenerational mobility in these data. We then employ BDR to estimate the joint conditional distribution of these two outcomes conditional on covariates for both the child and the father. We examine the role of the estimated local dependence parameter in generating these observed outcomes. Finally, we decompose the observed differences in dependence between fathers/sons and fathers/daughters into composition, sorting and marginals effects.[5]

6.1. Data. Our analysis examines data from the Panel Study of Income Dynamics (PSID), a rich longitudinal dataset including family and individual-level income data for individuals in the United States since 1968. The survey was conducted yearly until 1997 and biannually thereafter. The PSID has been employed in previous examiniations of intergenerational mobility (see, for example, Callaway and Huang (2019) or Landersø and Heckman (2017)).

[5]We acknowledge the use of the term "sorting" in this potential empirical context is somewhat inappropriate given that one would not interpret the process by which fathers and children are paired is best characterized by a sorting process. However, we retain this description to avoid the introduction of additional terminology.

Although the PSID data has 904,796 observations, the number for which we can construct father/children pairs is 6,6767 comprising 3,424 daughters and 3,243 sons. Our empirical analysis is based on this smaller sample. A pair is included if both the father and the child are observed at least once between the ages of 25 and 50. The outcome variables are their respective average annual labor earnings. This encompasses wages, business income, bonuses, and overtime payments. Labor earnings have been standardized to 1982 dollars. Note that we exclude observations with zero labor income earnings for that year. The average frequency of observations is the same for both father-daughter and father-son pairs. That is, 9 years for children and 15 years for fathers. The equal frequency for males and females is puzzling given the higher labor force participation rates of males. This might reflect the construction of the data set but probably also be due to our relatively weak requirement that the child only needs to be observed with earnings in one year. The higher average annual work hours for sons (2011) than daughters (1540) is consistent with males having a relatively higher level of labor market engagement than females.

The inclusion of only pairs for which both the father and child report labor income may introduce some form of selection bias. Moreover, not accounting for the frequency that the pair is observed may also introduce an additional form of bias. While each of these issues is interesting, we delay their treatment to future work as they are beyond the scope of this paper.

Studies in the literature rely on various income measures to study intergenerational mobility. These include, for example, labor earnings, tax records, or lifetime income imputed over the life cycle. Mogstad and Torsvik (2023) note that the estimated degree of mobility can vary considerably depending on the measure used. We focus on labor income, noting that intergenerational persistence tends to be stronger for broader income concepts (Landersø and Heckman, 2017).

The variables displayed in Table 1 are all available since 1968 with the exception of "college", which indicates that the individual has a college degree, which is available from 1975. Observations with missing entries for any of the variables are excluded. Table 1 indicates that the father's information is generally collected during the 1980s, while the children's information is typically gathered in the early 2000s. Fathers are observed later in life, mostly between the ages of 34 and 47, while children are observed earlier between the ages of 27 and 36. This also explains the large difference in labor income in favor of fathers. Unsurprisingly, we find higher levels of labor income for fathers with a college degree. The same is true for their children. This can be partially explained by the high level

Table 1. Descriptive Statistics PSID Data

		Education		Race	
Sons	All	No Coll.	College	Other	White
Labor Income Children (1000 USD)	17.84	16.76	22.09	13.58	19.98
Labor Income Father (1000 USD)	21.32	18.84	31.07	15.11	24.43
Age Children	31.71	31.93	30.86	31.51	31.81
Age Father	39.41	39.75	38.07	39.33	39.45
White Children	0.67	0.63	0.85	0.06	0.98
White Father	0.67	0.62	0.84	0.00	1.00
College Children	0.30	0.22	0.61	0.17	0.36
College Father	0.20	0.00	1.00	0.10	0.26
HH Size Children	2.58	2.65	2.32	2.52	2.61
HH Size Father	4.70	4.86	4.08	5.33	4.39
Children Origin Urb. 1	0.07	0.09	0.02	0.07	0.07
Children Origin Urb. 2	0.52	0.50	0.60	0.37	0.60
Children Origin Urb. 3	0.37	0.38	0.34	0.53	0.30
Children Origin Urb. 4	0.03	0.03	0.04	0.03	0.03
Decade Children	2004.63	2003.65	2008.51	2005.66	2004.12
Decade Parents	1983.55	1983.06	1985.48	1984.94	1982.86
Daughters					
Labor Income Children (1000 USD)	12.18	11.04	16.95	10.72	13.10
Labor Income Father (1000 USD)	21.14	18.78	30.99	14.99	24.98
Age Children	32.00	32.20	31.17	32.16	31.90
Age Father	39.37	39.69	38.02	39.75	39.13
White Children	0.63	0.57	0.85	0.06	0.98
White Father	0.62	0.56	0.84	0.00	1.00
College Children	0.37	0.29	0.68	0.24	0.45
College Father	0.19	0.00	1.00	0.08	0.27
HH Size Children	2.93	3.02	2.53	2.98	2.89
HH Size Father	4.79	4.95	4.14	5.52	4.34
Children Origin Urb. 1	0.06	0.06	0.03	0.07	0.05
Children Origin Urb. 2	0.55	0.54	0.57	0.39	0.64
Children Origin Urb. 3	0.37	0.38	0.35	0.52	0.28
Children Origin Urb. 4	0.03	0.02	0.05	0.02	0.03
Decade Children	2005.60	2004.60	2009.76	2005.57	2005.61
Decade Parents	1984.04	1983.51	1986.27	1984.31	1983.88

Notes: Number of observations 3426 for daughters and 3255 for sons. All statistics are means. Child origin refers to a categorical variable with 4 distinct values measuring the degree of urbanization where the child grew up.

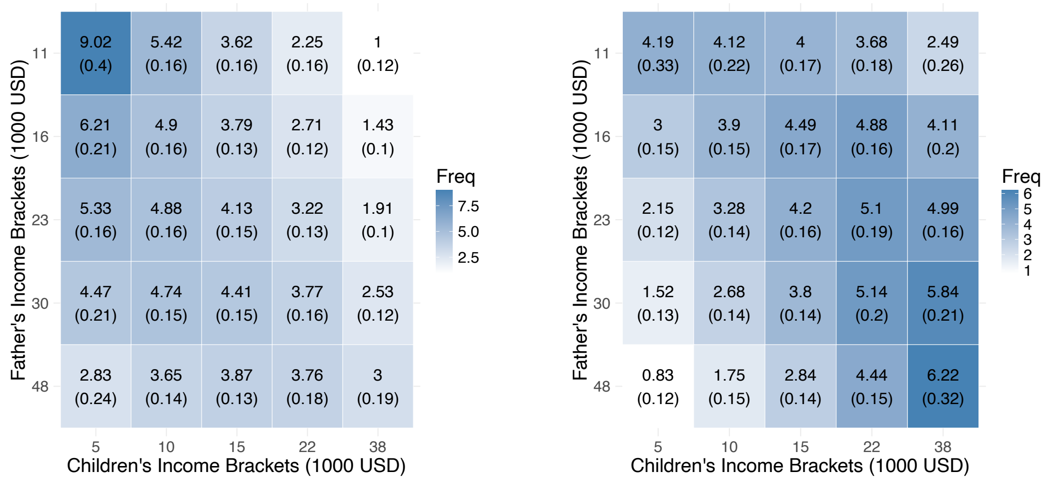
of persistence in college education. For example, 30 percent of sons have a college degree, but that number rises to 61 percent among sons whose fathers have a college degree. Similar figures are found for daughters. We also find that white fathers, and their children, generally have higher labor income. Household size is smaller for children, which can be attributed to their younger age, although it may also capture broader demographic changes and trends.

6.2. Unconditional Mobility. Figure 1 plots the transition matrices for the outcome variables. The cells of the transition matrices are based on the quintiles of the fathers in USD on the y-axis and the quintiles of the children in USD on the x-axis. The values within these cells represent the percentage of observations observed with a father-daughter or father-son combination within these cells. Due to the overall wage gap between women and men, sons earn more than daughters and the percentages of sons are relatively higher on the right side of the figure, while they are relatively higher on the left side of the figure for daughters. As relative darkness of the squares capture larger probabilities, it is immediately apparent that father-daughter pairs are relatively more frequently located in the upper left corner and the father-son pairs are more relatively frequent in the lower right corner. A high degree of mobility would imply equal probabilities across all cells within each row. Therefore, the observed deviation from this pattern illustrates the limited mobility in the United States.

6.3. BDR Estimation. We now focus on the estimates of the BDR model. We first estimate the model using the covariates described below for each of the marginals while also allowing ρ to be a function of the covariates. We also estimate the model in which ρ is not a function of the covariates although these latter results are relegated to the appendix.

The empirical results reported in the main text uses the following variables. For the marginal distribution of father's labor income we employ age, age^2 and decade fixed effects for both fathers and children, father's and child's race, father's and child's household size and indicators for a college degree of the father and the child's origin (measuring the degree of urbanization where the child was raised). We use the same set of variables for the marginal distribution of the children. When ρ is a function of the covariates we employ all of the covariates listed above but exclude the decade fixed effects. We acknowledge that some covariates are potentially endogenous. Accounting for this endogeneity is beyond the scope of this paper and we defer this important issue to further work. However, in the appendix we provide the corresponding results when we exclude some variables from the

FIGURE 1. Transition matrices, raw data (bootstrapped standard errors are shown in parentheses.)



(A) Daughters (B) Sons

Notes: The results in Figure 1 show the probability mass function for the raw data separately for sons and daughters. Reported standard errors equal the interquartile range of the bootstrap estimates divided by the interquartile range of the standard normal distribution.

conditioning set. We note that their exclusion does not have substantial implication for the results reported.

6.4. Model predictions and Counter Factual. We first report the capacity of the BDR estimates to reproduce the observed joint distribution of fathers' and children's earnings and examine the impact on this predicted joint distribution when we eliminate the dependence operating through ρ. Figure 2 presents the estimated transition matrices based on the BDR model. To construct these matrices, we estimate the joint distribution of father-child earnings using the set of covariates listed above. We then integrate over the empirical distribution of the observed covariates to obtain the average joint distribution.

Panels (A) and (B) display the resulting transition matrices for daughters and sons, respectively. These matrices correspond to the estimated joint distribution. The values of these matrices are calculated based on equation (3.4). Panels (C) and (D) show the same average distribution, but under the hypothetical assumption that the local correlation is set to zero. This counterfactual allows us to isolate the contribution of local dependence to intergenerational mobility. Panels (E) and (F) present the difference between the two distributions and quantifies the effect of local dependence.

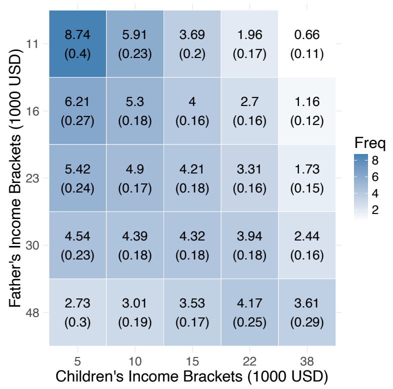
It is useful to consider two issues before focusing on the results from this exercise. The first is the interpretation of ρ in this context and the second is the resulting estimates of ρ. Given the nature of the outcomes, the parameter ρ captures how the outcomes move together after conditioning out the impact of the covariates. If the estimated value of ρ was zero, this implies no dependence between the outcomes after conditioning out the covariates. If the estimated ρ is positive (negative) this implies that, conditional on the covariates, a higher value for one outcome is associated with a higher (lower) outcome for the other. In this particular context this dependence may capture some local correlation between unobservables affecting the respective outcomes.

As noted above an attractive feature of BDR is its capacity to flexibly model the local dependence. Although we do not report our estimates of ρ, a notable feature of the results is the substantial heterogeneity in these estimates across different points on the grid. This reflects that the level of dependence varies greatly at different points of the joint distribution. This highlights the importance of adopting a flexible approach to its estimation.

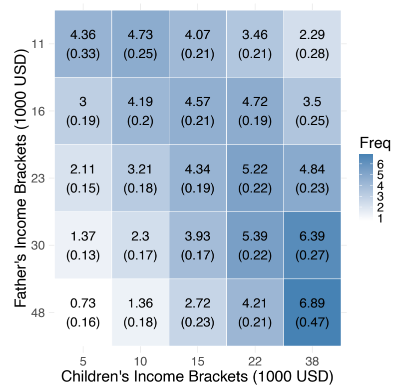
Several insights emerge from these results. First, the estimated transition matrices in Panels (A) and (B) closely resemble those obtained from the raw data (Figure 1), indicating that the model successfully captures the observed patterns of mobility. While this is reassuring it is not surprising given the nature of the estimator and the large number of parameters estimated. Second, setting the local correlation parameter (ρ) to zero leads to a considerable flattening of the dependence structure. This effect is evident for both sons and daughters and is particularly pronounced in the upper tail of the income distribution, where the level of dependence is relatively strong in the raw data. The influence of local dependence is quantitatively meaningful. For example, the probability that a son reaches the top decile is 6.9% under local dependence, compared to 5.8% under the counterfactual with $\rho = 0$. This difference is economically significant and explains a large fraction of the departure from the benchmark probability of 4%, which would be expected under full independence and in the absence of gender wage gaps and compositional effects. This highlights the role of local correlation in capturing some aspect of the observed persistent intergenerational advantages. Comparable patterns are observed for daughters, for whom the impact of ρ is most pronounced in the upper-left region of the figure.

Note that we also estimated the model predictions and the counterfactuals with a specification that does not take account of the covariates in ρ. These results are presented in Figure 4. The results are very similar to the results presented in Figure 2. This suggests that the results in this particulalr context are not sensitive to the treatment of the local dependence.

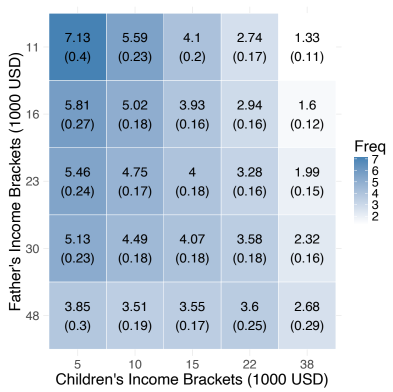
FIGURE 2. Counterfactual distributions (bootstrapped standard errors are shown in parentheses.)



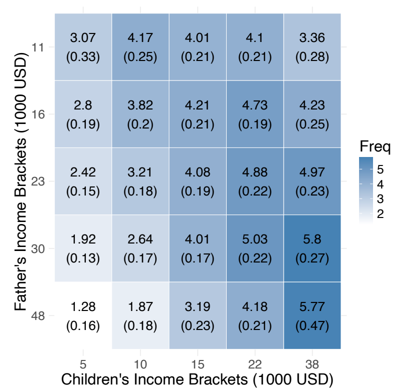
(A) Daughters



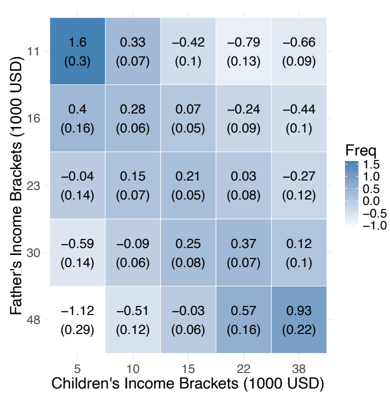
(B) Sons



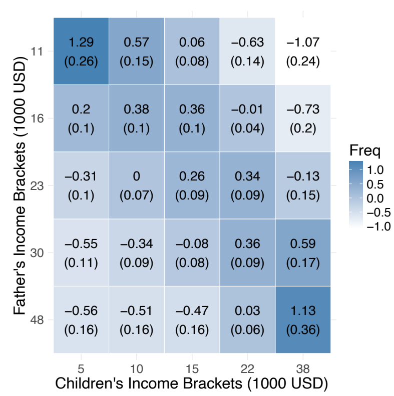
(C) Daughters, $\rho = 0$



(D) Sons, $\rho = 0$



(E) Daughters, difference due to ρ



(F) Sons, difference due to ρ

Notes: Panel (A) and (B) show the predicted probability mass functions from the full model. For panels (C) and (D), $\widehat{\rho}$ is artificially set to zero. The probability mass functions capture the dependence due to the marginals. Panel (E) and (F) show the differences between (A) and (C) and (B) and (D). Standard errors calculated as for Figure 1.

6.5. Decomposition of Joint Conditional CDF. Figure 3 presents a decomposition of the differences in transition matrices between sons and daughters. The decomposition follows equation (3.2) and separates the overall difference in the transition matrices into three distinct components: (i) Composition effects, which capture differences in observable characteristics (e.g. education, region, cohort); (ii) sorting effects, reflecting differences in the estimated local dependence parameter ρ; and (iii) marginal effects, which account for gender-specific differences in how covariates relate to the marginal earnings distributions.

Panels (A) and (B) display the estimated transition matrices for sons and daughters, respectively, using the BDR model with the set of covariates as described in Section 6.3. Panel (C) shows the element-wise difference between the two matrices. Panels (D), (E), and (F) then report how much of this difference can be explained by composition, sorting, and marginal effects, respectively, expressed as a percentage of the absolute difference in each cell.[6]

Several key insights emerge from this analysis. Marginal effects are by far the most important driver of the observed gender differences in intergenerational earnings mobility. In many cells, they account for the majority of the variation across gender, highlighting the importance of differential returns to characteristics in shaping the marginal earnings distributions of sons and daughters. These marginal effects capture the differences in the labor market experiences, reflecting both supply and demand factors, for females and males. Sorting effects, driven by differences in the estimated local correlation, explain roughly 10% of the overall difference. This suggests that sons and daughters differ not only in their marginal earnings prospects, but also in how tightly their outcomes are linked to their fathers' incomes, even after controlling for observed characteristics. Composition effects contribute very little to the observed differences. This indicates that the sons and daughters in our sample are, on average, very similar in terms of the included covariates. This suggests that the large difference in intergenerational earnings mobility for daughters and sons cannot be explained by differences in observable characteristics. These results emphasize that gender differences in intergenerational earnings mobility arise primarily from structural differences in earnings determination. These arise through the impact of observables on earnings and the nature of dependence between father's and child's outcomes.

We also provide a robustness check that excludes the children's race and household size. The results are presented in Table 5. We find no substantial deviations from the

[6]Note that the difference in Panel (C) is, in some instances, very close to zero, which means that the composition, sorting, and marginal effects are divided by a number near zero. This partially explains the occasional occurrence of very large values in Panels (D)–(F).

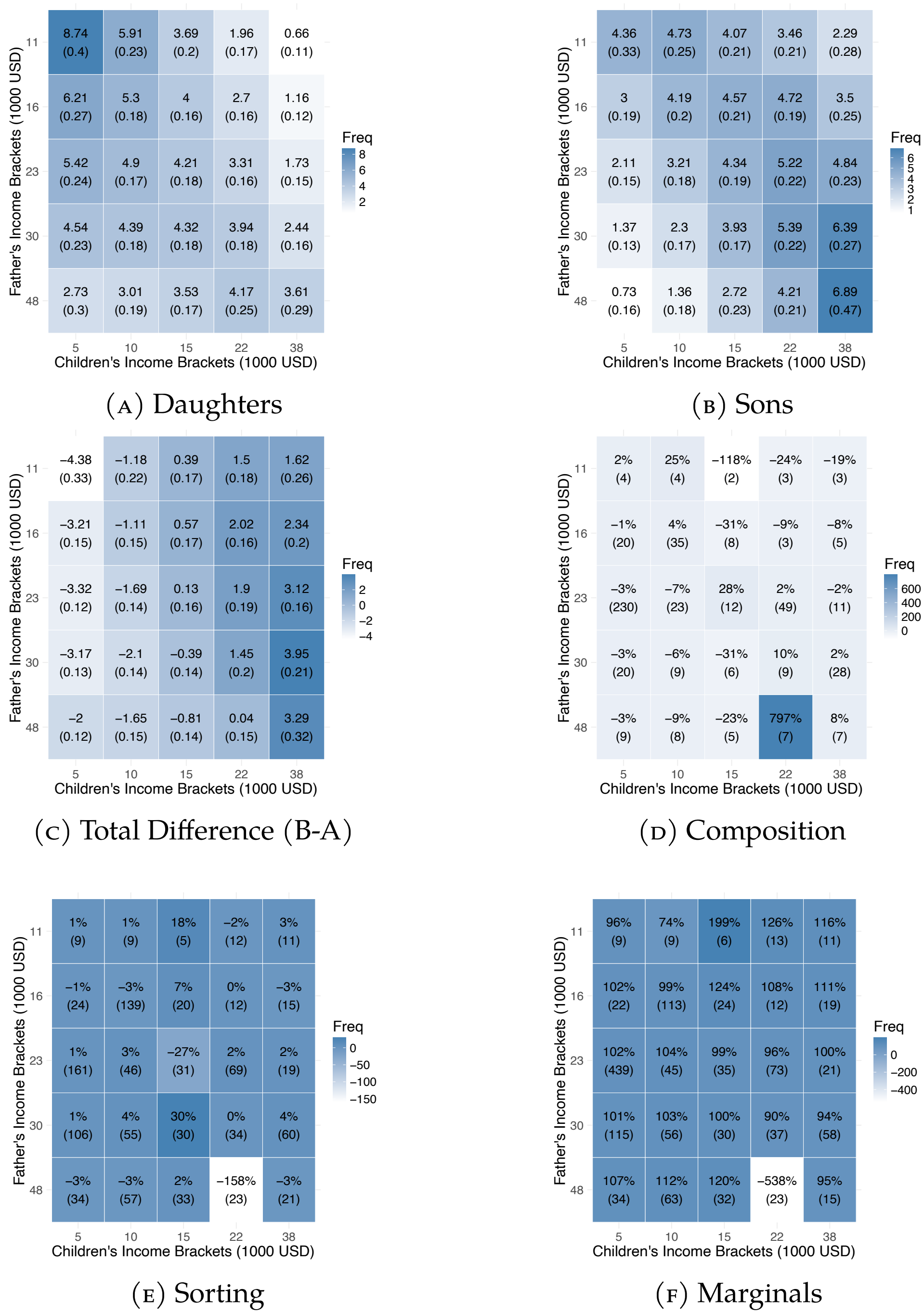
results shown in Figure 3. While we do not consider this an exhaustive examination of the sensitivity of our results, it does not provide some indication of their sensitivity to the exclusion of variables which are potentially important.

6.6. Discussion. Our empirical findings highlight the importance of modeling both the marginal distributions and the local dependence structure when analyzing intergenerational mobility. The BDR framework enables a flexible and interpretable decomposition of mobility, and highlights that both observable characteristics and residual local association patterns contribute to the overall structure. Importantly, one can isolate the channels through which gender differences in income dependence emerge. The stronger association observed in father-son pairs is largely due to the differences in the marginal distributions. However, accounting for differences in sorting, capturing local dependence, is also important. Overall, the evidence suggests that policies aiming to promote intergenerational equity should consider not just differences in endowments, but also structural differences in how income is transmitted across generations.

7. Conclusion

We employ bivariate distribution regression to estimate the joint distribution of two outcome variables conditional on set(s) of covariates. Our approach is particularly valuable when the dependence between the outcomes persists even after controlling for observed covariates. We describe how BDR can be implemented and present some associated functionals of interest. The functionals of primary interest in this setting are those involving the unexplained dependence. We provide a decomposition of the joint distributions for different groups into composition, marginal and sorting effects. We provide algorithms for estimation and inference, along with the associated theory. We also provide a similar decomposition for the associated transition matrices Using data from the Panel Survey of Income Dynamics, we model the joint distribution of parents' and children's earnings and find that the local dependence structure is an important contributor in explaining intergenerational mobility.

Figure 3. Decomposition of difference in transition matrices between sons and daughters (bootstrapped standard errors in parentheses).



Notes: Figure 3 shows the decomposition of the joint distribution between sons and daughters into three terms. Values in Panels (D), (E), and (F) are in percentage differences of the total difference. Note that there is not an equal number of observations in all rows as sons and daughters have different marginal distributions. For this figure, we evaluated the joint at the common labor income levels. Standard errors calculated as in Figure 1.

Appendix A. Proofs

A.1. Notation and Auxiliary Result for Lemma 5.1. Let

$$\mathbb{G}_n[f] := \mathbb{G}_n[f(W)] := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(W) - \mathrm{E}[f(W)]).$$

When the function $\widehat{f}$ is estimated, the notation should be interpreted as

$$\mathbb{G}_n[\widehat{f}] = \mathbb{G}_n[f]|_{f=\widehat{f}}.$$

We use the Z-process framework from Appendix E.1 of Chernozhukov et al. (2013). The score vector in (5.3) has the analytical form:

$$S_i^{yw}(\theta_{yw}^d) =$$
$$\begin{bmatrix} \dfrac{\phi(X_i^{d'}\mu_y^d)}{\Phi(X_i^{d'}\mu_y^d)\Phi(-X_i^{d'}\mu_y^d)}\left[I_i^{dy} - \Phi(X_i^{d'}\mu_y^d)\right] \\ \dfrac{\phi(X_i^{d'}\nu_w^d)}{\Phi(X_i^{d'}\nu_w^d)\Phi(-X_i^{d'}\nu_w^d)}\left[J_i^{dw} - \Phi(X_i^{d'}\nu_w^d)\right] \\ \left(\dfrac{I_i^{dy}J_i^{dw}}{\Phi_{2i}^d(y,w)} - \dfrac{I_i^{dy}\bar{J}_i^{dw}}{\Phi_{2i}^d(y,-w)} - \dfrac{\bar{I}_i^{dy}J_i^{w}}{\Phi_{2i}^d(-y,w)} + \dfrac{\bar{I}_i^{dy}\bar{J}_i^{dw}}{\Phi_{2i}^d(-y,-w)}\right)\phi_2(X_i^{d'}\mu_y, X_i^{d'}\nu_w; g(X_i^{d'}\delta_{yw}^d))\dot{g}(X_i^{d'}\delta_{yw}^d) \end{bmatrix} \otimes X_i^d, \tag{A.1}$$

where $\Phi_{2i}^d(y,w) := \Phi_2(X_i^{d'}\mu_y^d, X_i^{d'}\nu_w^d; g(X_i^{d'}\delta_{yw}^d))$, $\Phi_{2i}^d(y,-w) := \Phi_2(X_i^{d'}\mu_y^d, -X_i^{d'}\nu_w^d; -g(X_i^{d'}\delta_{yw}^d))$, $\Phi_{2i}^d(-y,w) := \Phi_2(-X_i^{d'}\mu_y^d, X_i^{d'}\nu_w^d; -g(X_i^{d'}\delta_{yw}^d))$, $\Phi_{2i}^d(-y,-w) := \Phi_2(-X_i^{d'}\mu_y^d, -X_i^{d'}\nu_w^d; g(X_i^{d'}\delta_{yw}^d))$, and $\dot{g}(u) := \mathrm{d}g(u)/\mathrm{d}u$. The non-zero components of the expected Hessian matrix in (5.4) have the analytical form

$$\mathrm{E}\left[\partial_{\mu\mu}\ell_i^y(\mu_y^d)\right] = -\mathrm{E}\left[\frac{\phi(X_i^{d'}\mu_y^d)^2}{\Phi(X_i^{d'}\mu_y^d)\Phi(-X_i^{d'}\mu_y^d)}X_i^d X_i^{d'}\right],$$
$$\mathrm{E}\left[\partial_{\nu\nu}\ell_i^w(\nu_w^d)\right] = -\mathrm{E}\left[\frac{\phi(X_i^{d'}\nu_w^d)^2}{\Phi(X_i^{d'}\nu_w^d)\Phi(-X_i^{d'}\nu_y^d)}X_i^d X_i^{d'}\right],$$

$$\mathrm{E}\left[\partial_{\delta\mu}\ell_i^{yw}(\theta_{yw}^d)\right]$$
$$= -\mathrm{E}\left[\left\{\frac{\Phi_{2i}^{d\mu}(y,w)}{\Phi_{2i}^d(y,w)} - \frac{\Phi_{2i}^{d\mu}(y,-w)}{\Phi_{2i}^d(y,-w)} + \frac{\Phi_{2i}^{d\mu}(y,w)}{\Phi_{2i}^d(-y,w)} - \frac{\Phi_{2i}^{d\mu}(y,-w)}{\Phi_{2i}^d(-y,-w)}\right\}\Phi_{2i}^{d\rho}(y,w)\dot{g}(X^{d'}\delta_{yw}^d)X_i^d X_i^{d'}\right],$$

$$\mathrm{E}\left[\partial_{\delta\nu}\ell_i^{yw}(\theta_{yw}^d)\right]$$
$$= -\mathrm{E}\left[\left\{\frac{\Phi_{2i}^{d\nu}(y,w)}{\Phi_{2i}^d(y,w)} + \frac{\Phi_{2i}^{d\nu}(y,w)}{\Phi_{2i}^d(y,-w)} - \frac{\Phi_{2i}^{d\nu}(-y,w)}{\Phi_{2i}^d(-y,w)} - \frac{\Phi_{2i}^{d\nu}(-y,w)}{\Phi_{2i}^d(-y,-w)}\right\}\Phi_{2i}^{d\rho}(y,w)\dot{g}(X_i^{d'}\delta_{yw}^d)X_i^d X_i^{d'}\right],$$

$$
\begin{aligned}
&\mathrm{E}\left[\partial_{\delta\delta}\ell_i^{yw}(\theta_{yw}^d)\right] \\
&\quad = -\mathrm{E}\left[\left\{\frac{1}{\Phi_{2i}^d(y,w)} + \frac{1}{\Phi_{2i}^d(y,-w)} + \frac{1}{\Phi_{2i}^d(-y,w)} + \frac{1}{\Phi_{2i}^d(-y,-w)}\right\}\Phi_{2i}^{d\rho}(y,w)^2\dot{g}(X_i^{d'}\delta_{yw}^d)^2 X_i^d X_i^{d'}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
\Phi_{2i}^{d\mu}(\pm y, \mp w) &= \partial_\mu\ \Phi_2(\pm X_i^{d'}\mu_y^d, \mp X_i^{d'}\nu_w^d; g(X_i^{d'}\delta_{yw}^d))\Big|_{\mu=\pm X_i^{d'}\mu_y^d} \\
&= \Phi\left(\frac{\mp X_i^{d'}\nu_w^d - g(X_i^{d'}\delta_{yw}^d)(\pm X_i^{d'}\mu_y^d)}{\sqrt{1-g(X_i^{d'}\delta_{yw}^d)^2}}\right)\phi(X_i^{d'}\mu_y^d),
\end{aligned}
$$

$$
\begin{aligned}
\Phi_{2i}^{d\nu}(\pm y, \mp w) &= \partial_\nu\ \Phi_2(\pm X_i^{d'}\mu_y^d, \mp X_i^{d'}\nu_w^d; g(X_i^{d'}\delta_{yw}^d))\Big|_{\nu=X_i^{d'}\nu_w^d} \\
&= \Phi\left(\frac{\pm X_i^{d'}\mu_y^d - g(X_i^{d'}\delta_{yw}^d)(\mp X_i^{d'}\nu_w^d)}{\sqrt{1-g(X_i^{d'}\delta_{yw}^d)^2}}\right)\phi(X_i^{d'}\nu_w^d),
\end{aligned}
$$

and

$$
\Phi_{2i}^{d\rho}(y,w) = \partial_\rho\ \Phi_2(X_i^{d'}\mu_y^d, X_i^{d'}\nu_w^d; \rho)\Big|_{\rho=g(X_i^{d'}\delta_{yw}^d)} = \phi_2(X_i^{d'}\mu_y^d, X_i^{d'}\nu_w^d; g(X_i^{d'}\delta_{yw}^d)).
$$

Also, by the inverse of the partitioned matrix formula, the inverse of $H(\theta_{yw}^d)$ is

$$
\begin{bmatrix}
\mathrm{E}[\partial_{\mu\mu}\ell_i^y(\mu_y^d)]^{-1} & 0 & 0 \\
0 & \mathrm{E}[\partial_{\nu\nu}\ell_i^w(\nu_w^d)]^{-1} & 0 \\
-\mathrm{E}[\partial_{\delta\delta}\ell_i^{yw}(\theta_{yw}^d)]^{-1}\mathrm{E}[\partial_{\delta\mu}\ell_i^{yw}(\theta_{yw}^d)]\mathrm{E}[\partial_{\mu\mu}\ell_i^y(\mu_y^d)]^{-1} & -\mathrm{E}\partial_{\delta\delta}\ell_i^{yw}(\theta_{yw}^d)]^{-1}\mathrm{E}\partial_{\delta\nu}\ell_i^{yw}(\theta_{yw}^d)]\mathrm{E}[\partial_{\nu\nu}\ell_i^w(\nu_w^d)]^{-1} & \partial_{\delta\delta}\ell_i^{yw}(\theta_{yw}^d)^{-1}
\end{bmatrix}.
$$

The following lemma, which combines results from Chernozhukov et al. (2013) and Chernozhukov et al. (2018), provides sufficient conditions to verify Condition Z in Chernozhukov et al. (2013) on the compact set $\bar{\mathcal{Y}} \times \bar{\mathcal{W}}$. Let $(\theta, y, w) \mapsto \Psi(\theta, y, w) := \mathrm{E}[S_i^{yw}(\theta)]$ and $\Theta = \mathcal{M} \times \mathcal{N} \times \mathcal{D}$ be the parameter space for $\theta_{yw}^d = (\mu_y^{d'}, \nu_w^{d'}, \delta_{yw}^{d'})'$.

Lemma A.1 (Sufficient Condition for Z)**.** *Suppose that $\mathcal{M} = \mathcal{N} = \mathbb{R}^{d_x}$, $\mathcal{D}$ is a compact subset of $\mathbb{R}^{d_x}$ and $\bar{\mathcal{Y}}\bar{\mathcal{W}} := \bar{\mathcal{Y}} \times \bar{\mathcal{W}}$ is a compact set in $\mathbb{R}^2$. Let $\mathcal{I}$ be an open set containing $\bar{\mathcal{Y}}\bar{\mathcal{W}}$. Suppose that (a) $\Psi : \Theta \times \mathcal{I} \mapsto \mathbb{R}^{d_\theta}$ is continuous, (b) $\mu \mapsto \mathrm{E}[]\partial_\mu\ell_i^y(\mu)]$ and $\nu \mapsto \mathrm{E}[\partial_\nu\ell_i^w(\nu)]$ are the gradients of convex functions for each $y \in \bar{\mathcal{Y}}$ and $w \in \bar{\mathcal{W}}$, respectively, and $\delta \mapsto \mathrm{E}[\partial_\delta\ell_i^{yw}(\mu_y^d, \nu_w^d, \delta)]$ possesses a unique zero at δ_{yw} that is in the interior of $\mathcal{D}$ for each $(y,w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}}$ and $d \in \{0,1\}$, (c) $\partial\Psi(\theta, y, w)/\partial(\theta', y, w)$ exists at (θ_{yw}^d, y, w), and is continuous at (θ_{yw}^d, y, w) for each $(y,w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}}$ and $d \in \{0,1\}$, and $\dot{\Psi}_{\theta_{yw}^d,y,w} = \partial\Psi(\theta,y,w)/\partial\theta'|_{\theta_{yw}^d}$ obeys $\inf_{(y,w)\in\bar{\mathcal{Y}}\bar{\mathcal{W}}}\inf_{||h||=1}\left\|\dot{\Psi}_{\theta_{yw}^d,y,w}h\right\| >$*

$c_0 > 0$ *for each* $d \in \{0, 1\}$*. Then Condition Z of Chernozhukov et al. (2013) holds and* $(y, w) \mapsto \theta^d_{yw}$ *is continuously differentiable for each* $d \in \{0, 1\}$*.*

A.2. Proof of Lemma 5.1. We drop the group index d to lighten the notation whenever it does not cause confusion. An identical proof applies to both groups and the joint convergence follows from the marginal convergence by independence.

We first show the result of the unrestricted estimator $\widehat{\theta}_{yw}$ and then for the estimator of the the tail parameters.

FCLT and Bootstrap FCLT for $\widehat{\theta}^d_{yw}$. The proof follows the same steps as the proof of Theorem 5.2 of Chernozhukov et al. (2013) and Theorem C.1 of Chernozhukov et al. (2018). Let $\Psi(\theta, y, w) := \mathrm{P}\left[S^{yw}_i(\theta)\right]$ and $\widehat{\Psi}(\theta, y, w) := \mathrm{P}_n\left[S^{yw}_i(\theta)\right]$, where P_n represents the empirical measure and P represents the corresponding probability measure. From the first-order conditions, $\widehat{\theta}_{yw} = \phi\left(\widehat{\Psi}(\cdot, y, w), 0\right)$ for each $(y, w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}}$, where ϕ is the Z-map defined in Appendix E.1 of Chernozhukov et al. (2013). The random vector $\widehat{\theta}_{yw}$ is the estimator of $\theta_{yw} = \phi\left(\Psi(\cdot, y, w), 0\right)$ in the notation of this framework. Applying Step 1 below, we obtain

$$\sqrt{n}\left(\widehat{\Psi} - \Psi\right) \rightsquigarrow Z^{\Psi} \text{ in } \ell^{\infty}\left(\bar{\mathcal{Y}}\bar{\mathcal{W}} \times \mathbb{R}^{d_\theta}\right)^{d_\theta},$$

where $Z_{\Psi}(y, w, \theta) = \mathbb{G}S^{yw}(\theta)$, $\mathbb{G}$ is a P-Brownian bridge, and Z_{Ψ} has continuous paths a.s.

Step 2 verifies the conditions of Lemma A.1 for Ψ, which also implies that $(y, w) \mapsto \theta_{yw}$ is continuously differentiable in the set $\bar{\mathcal{Y}}\bar{\mathcal{W}}$. Then, by Lemma E.2 of Chernozhukov et al. (2013), the map ϕ is Hadamard differentiable with derivative map $(\psi, 0) \mapsto -\dot{\Psi}^{-1}\psi$ at $(\Psi, 0)$. Therefore, we can conclude that

$$\sqrt{n}\left(\widehat{\theta}^d_{yw} - \theta^d_{yw}\right) \rightsquigarrow Z^{d\theta}_{yw} := -\sqrt{p_d}\,\dot{\Psi}^{-1}_{\theta^d_{yw}, y, w} Z^{\Psi}(y, w, \theta^d_{yw}) \text{ in } \ell^{\infty}\left(\bar{\mathcal{Y}}\bar{\mathcal{W}}\right)^{d_\theta}, \tag{A.2}$$

where $(y, w) \mapsto Z^{d\theta}_{yw}$ has continuous paths a.s., because

$$\sqrt{n_d}\left(\widehat{\theta}^d_{yw} - \theta^d_{yw}\right) \rightsquigarrow -\dot{\Psi}^{-1}_{\theta^d_{yw}, y, w} Z^{\Psi}(y, w, \theta^d_{yw}) \text{ in } \ell^{\infty}\left(\bar{\mathcal{Y}}\bar{\mathcal{W}}\right)^{d_\theta},$$

by the functional delta method and

$$\sqrt{n_d}/\sqrt{n} \to_{\mathrm{P}} \sqrt{p_d}.$$

Step 1 (Donskerness)**.** We verify that $\mathcal{G} = \left\{S^{yw}_i(\theta) : (y, w, \theta) \in \bar{\mathcal{Y}}\bar{\mathcal{W}} \times \mathbb{R}^{d_\theta}\right\}$ is a P-Donsker with a square-integrable envelope. By examining the expression in (A.1), we see that $S^{yw}_i(\theta)$ is a Lipschitz transformation of VC functions with Lipschitz coefficient bounded by $c\,\|X_i\|$ for some constant c and envelope function $c\,\|X_i\|$, which is square-integrable. Hence $\mathcal{G}$ is P-Donsker by Example 19.9 in Vaart (2000).

Step 2 (Verification of the Conditions of Lemma A.1)**.** Conditions (a) and (b) follow directly from Assumption 5.1. To verify (c), note that for $(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w})$ in the neighborhood of (θ_{yw}, y, w),

$$\frac{\partial \Psi(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w})}{\partial(\tilde{\theta}'_{yw}, \tilde{y}, \tilde{w})} = \left[H^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}), J^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}), K^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) \right].$$

To characterize the analytical expressions of $H^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw})$, $J^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw})$ and $K^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw})$, we use the notation and expressions introduced in Appendix A.1. Let $\tilde{\Phi}_{2i}(y,w)$, $\tilde{\Phi}_{2i}(y,-w)$, $\tilde{\Phi}_{2i}(-y,w)$, $\tilde{\Phi}_{2i}(-y,-w))$, $\tilde{\Phi}^{\mu}_{2i}(\pm y, \mp w)$, $\tilde{\Phi}^{\nu}_{2i}(\pm y, \mp w)$ and $\tilde{\Phi}^{\rho}_{2i}(y,w)$ be the same expressions as $\Phi_{2i}(y,w)$, $\Phi_{2i}(y,-w)$, $\Phi_{2i}(-y,w)$, $\Phi_{2i}(-y,-w))$, $\Phi^{\mu}_{2i}(\pm y, \mp w)$, $\Phi^{\nu}_{2i}(\pm y, \mp w)$ and $\Phi^{\rho}_{2i}(y,w)$ in Appendix A.1 but evaluated at $\tilde{\theta}_{yw} = (\tilde{\mu}_y, \tilde{\nu}_w, \tilde{\delta}_{yw})$.[7] Then,

$$J^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) := \frac{\partial \Psi(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w})}{\partial \tilde{y}} = \mathrm{E}\begin{bmatrix} \dfrac{\phi(X_i'\tilde{\mu}_y)}{\Phi(X_i'\tilde{\mu}_y)\Phi(-X_i'\tilde{\mu}_y)} f_{Y|X}(\tilde{y} \mid X) \\ 0 \\ J_3^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) \end{bmatrix},$$

where

$$J_3^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = \left(\frac{1}{\tilde{\Phi}_{2i}(y,-w)} - \frac{1}{\tilde{\Phi}_{2i}(-y,-w)} \right) f_{Y|X}(\tilde{y} \mid X) \tilde{\Phi}^{\rho}_{2i}(y,w) \dot{g}(X_i'\tilde{\delta}_{yw}) X_i$$
$$+ \left(\frac{1}{\tilde{\Phi}_{2i}(y,w)} - \frac{1}{\tilde{\Phi}_{2i}(y,-w)} - \frac{1}{\tilde{\Phi}_{2i}(-y,w)} + \frac{1}{\tilde{\Phi}_{2i}(-y,-w)} \right) F_{Y|W,X}(\tilde{y} \mid \tilde{w}, X) f_{W|X}(\tilde{w} \mid X) \tilde{\Phi}^{\rho}_{2i}(y,w) \dot{g}(X_i'\tilde{\delta}_{yw}) X_i;$$

and

$$K^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) := \frac{\partial \Psi(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w})}{\partial \tilde{w}} = \mathrm{E}\begin{bmatrix} 0 \\ \dfrac{\phi(X_i'\tilde{\nu}_w)}{\Phi(X_i'\tilde{\nu}_w)\Phi(-X_i'\tilde{\nu}_w)} f_{W|X}(\tilde{w} \mid X) \\ K_3^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) \end{bmatrix},$$

where

$$K_3^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = \left(\frac{1}{\tilde{\Phi}_{2i}(-y,w)} - \frac{1}{\tilde{\Phi}_{2i}(-y,-w)} \right) f_{W|X}(\tilde{w} \mid X) \tilde{\Phi}^{\rho}_{2i}(y,w) \dot{g}(X_i'\tilde{\delta}_{yw}) X_i$$
$$+ \left(\frac{1}{\tilde{\Phi}_{2i}(y,w)} - \frac{1}{\tilde{\Phi}_{2i}(y,-w)} - \frac{1}{\tilde{\Phi}_{2i}(-y,w)} + \frac{1}{\tilde{\Phi}_{2i}(-y,-w)} \right) F_{W|Y,X}(\tilde{w} \mid \tilde{y}, X) f_{Y|X}(\tilde{y} \mid X) \tilde{\Phi}^{\rho}_{2i}(y,w) \dot{g}(X_i'\tilde{\delta}_{yw}) X_i.$$

Also,

$$H^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) := \frac{\partial \Psi(\tilde{y}, \tilde{s}, \tilde{\theta}_{yw})}{\partial \tilde{\theta}'_{yw}} = \mathrm{E}\begin{bmatrix} H_{11}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) & 0 & 0 \\ 0 & H_{22}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) & 0 \\ H_{31}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) & H_{32}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) & H_{33}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) \end{bmatrix}.$$

[7]Note that $\tilde{\Phi}_{2i}(y,w)$ is different from $\Phi_{2i}(\tilde{y},\tilde{w})$, because the former is evaluated at $\tilde{\theta}_{yw}$, a value near θ_{yw}, whereas the latter is evaluated at $\theta_{\tilde{y}\tilde{w}}$, the true value of θ_{yw} at $(y,w) = (\tilde{y},\tilde{w})$.

where

$$H_{11}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = \left\{ G(X_i'\tilde{\mu}_y)\left[\Phi(X_i'\mu_{\tilde{y}}) - \Phi(X_i'\tilde{\mu}_y)\right] - \frac{\phi(X_i'\mu_y)^2}{\Phi(X_i'\mu_y)\Phi(-X_i'\mu_y)} \right\} X_i X_i',$$

$$H_{22}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = \left\{ G(X_i'\tilde{\nu}_w)\left[\Phi(X_i'\nu_{\tilde{w}}) - \Phi(X_i'\tilde{\nu}_w)\right] - \frac{\phi(X_i'\nu_w)^2}{\Phi(X_i'\nu_w)\Phi(-X_i'\nu_w)} \right\} X_i X_i',$$

$$\begin{aligned} H_{31}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = & \\ -\Bigg\{ & \frac{\tilde{\Phi}_{2i}^{\mu}(y,w)\Phi_{2i}(\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(y,w)^2} - \frac{\tilde{\Phi}_{2i}^{\mu}(y,-w)\Phi_{2i}(\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(y,-w)^2} + \frac{\tilde{\Phi}_{2i}^{\mu}(y,w)\Phi_{2i}(-\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(-y,w)^2} - \frac{\tilde{\Phi}_{2i}^{\mu}(y,-w)\Phi_{2i}(-\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(-y,-w)^2} \Bigg\} \times \\ & \tilde{\Phi}_{2i}^{\rho}(y,w)\dot{g}(X_i'\tilde{\delta}_{yw})X_iX_i' \\ & - \left\{ \frac{\Phi_{2i}(\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(y,w)} - \frac{\Phi_{2i}(\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(y,-w)} - \frac{\Phi_{2i}(-\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(-y,w)} + \frac{\Phi_{2i}(-\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(-y,-w)} \right\} \tilde{\Phi}_{2i}^{\rho\mu}(y,w)\dot{g}(X_i'\tilde{\delta}_{yw})X_iX_i', \end{aligned}$$

$$\begin{aligned} H_{32}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = & -\left\{ \frac{\Phi_{2i}(\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(y,w)} - \frac{\Phi_{2i}(\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(y,-w)} - \frac{\Phi_{2i}(-\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(-y,w)} + \frac{\Phi_{2i}(-\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(-y,-w)} \right\} \tilde{\Phi}_{2i}^{\rho\nu}(y,w)\dot{g}(X_i'\tilde{\delta}_{yw})X_iX_i' \\ -\Bigg\{ & \frac{\tilde{\Phi}_{2i}^{\nu}(y,w)\Phi_{2i}(\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(y,w)^2} + \frac{\tilde{\Phi}_{2i}^{\nu}(y,w)\Phi_{2i}(\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(y,-w)^2} - \frac{\tilde{\Phi}_{2i}^{\nu}(-y,w)\Phi_{2i}(-\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(-y,w)^2} - \frac{\tilde{\Phi}_{2i}^{\nu}(-y,w)\Phi_{2i}(-\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(-y,-w)^2} \Bigg\} \times \\ & \tilde{\Phi}_{2i}^{\rho}(y,w)\dot{g}(X_i'\tilde{\delta}_{yw})X_iX_i', \end{aligned}$$

$$\begin{aligned} H_{33}^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw}) = & -\left\{ \frac{\Phi_{2i}(\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(y,w)} - \frac{\Phi_{2i}(\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(y,-w)} - \frac{\Phi_{2i}(-\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(-y,w)} + \frac{\Phi_{2i}(-\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(-y,-w)} \right\} \times \\ & \left[\tilde{\Phi}_{2i}^{\rho\rho}(y,w)\dot{g}(X_i'\tilde{\delta}_{yw}) + \tilde{\Phi}_{2i}^{\rho}(y,w)\ddot{g}(X_i'\tilde{\delta}_{yw}) \right] X_iX_i' \\ & - \left\{ \frac{\Phi_{2i}(\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(y,w)^2} + \frac{\Phi_{2i}(\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(y,-w)^2} + \frac{\Phi_{2i}(-\tilde{y},\tilde{w})}{\tilde{\Phi}_{2i}(-y,w)^2} + \frac{\Phi_{2i}(-\tilde{y},-\tilde{w})}{\tilde{\Phi}_{2i}(-y,-w)^2} \right\} \tilde{\Phi}_{2i}^{\rho}(y,w)^2\dot{g}(X_i'\tilde{\delta}_{yw})X_iX_i', \end{aligned}$$

$$G(u) = \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{\phi(u)}{\Phi(u)\Phi(-u)}\right] = -\frac{\phi(u)}{\Phi(u)\Phi(-u)}\left[u + \frac{1-2\Phi(u)}{\Phi(u)\Phi(-u)}\right],$$

$$\tilde{\Phi}_{2i}^{\rho\mu} = \partial^2_{\rho,\mu}\ \Phi_2(\mu, X_i'\tilde{\nu}_w;\rho)\big|_{\mu=X_i'\tilde{\mu}_y,\rho=g(X_i'\tilde{\delta}_{yw})} = -\frac{X_i'\tilde{\mu}_y - g(X_i'\tilde{\delta}_{yw})X_i'\tilde{\nu}_w}{1-g(X_i'\tilde{\delta}_{yw})^2}\phi_2(X_i'\tilde{\mu}_y, X_i'\tilde{\nu}_w; g(X_i'\tilde{\delta}_{yw})),$$

$$\tilde{\Phi}_{2i}^{\rho\nu} = \partial^2_{\rho,\nu}\ \Phi_2(X_i'\tilde{\mu}_y, \nu;\rho)\big|_{\nu=X_i'\tilde{\nu}_w,\rho=g(X_i'\tilde{\delta}_{yw})} = -\frac{X_i'\tilde{\nu}_w - g(X_i'\tilde{\delta}_{yw})X_i'\tilde{\mu}_y}{1-g(X_i'\tilde{\delta}_{yw})^2}\phi_2(X_i'\tilde{\mu}_y, X_i'\tilde{\nu}_w; g(X_i'\tilde{\delta}_{yw})),$$

and

$$\tilde{\Phi}_{2i}^{\rho\rho} = \partial^2_{\rho,\rho}\ \Phi_2(X_i'\tilde{\mu}_y, X_i'\tilde{\nu}_w;\rho)\big|_{\rho=g(X_i'\tilde{\delta}_{yw})} = \partial_\rho\ \phi_2(X_i'\tilde{\mu}_y, X_i'\tilde{\nu}_w;\rho)\big|_{\rho=g(X_i'\tilde{\delta}_{yw})}.$$

We want to show that $(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w}) \mapsto H^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw})$, $(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w}) \mapsto J^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw})$, and $(\tilde{\theta}_{yw}, \tilde{y}, \tilde{w}) \mapsto K^{\tilde{y}\tilde{w}}(\tilde{\theta}_{yw})$ are continuous at (θ_{yw}, y, w) for each $(y,w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}}$. The verification of the continuity follows from using the expressions, the dominated convergence theorem, and the following ingredients: (1) a.s. continuity of the map $(\tilde\theta_{yw}, \tilde s, \tilde y) \mapsto \partial_\theta S_i^{\tilde y \tilde w}(\tilde\theta_{yw})$, (2) domination of $\left\|\partial_\theta S_i^{\tilde y \tilde w}(\tilde\theta_{yw})\right\|$ by a square-integrable function $c\,\|X_i\|$ for some constant $c > 0$, (3) a.s. continuity and uniform boundedness of the conditional density functions $y \mapsto f_{Y|X}(y \mid x)$ and $w \mapsto f_{W|X}(w \mid x)$ by Assumption 5.1, and (4) $\tilde\Phi_{2i}(y,w)$, $\tilde\Phi_{2i}(y,-w)$, $\tilde\Phi_{2i}(-y,w)$, and $\tilde\Phi_{2i}(-y,-w))$ being bounded away from zero, and $\tilde\Phi^{\mu}_{2i}(\pm y, \mp w)$, $\tilde\Phi^{\nu}_{2i}(\pm y, \mp w)$ and $\tilde\Phi^{\rho}_{2i}(y,w)$ being bounded, uniformly on $\tilde\theta_{yw} \in \mathbb{R}^{d_\theta}$ a.s. Finally, the last part of (c) follows because the minimum eigenvalue of $\dot\Psi_{\theta_{yw},y,w} = H^{yw}(\theta_{wy})$ is bounded away from zero uniformly on $(y,w) \in \bar{\mathcal{Y}}\bar{\mathcal{W}}$ by Assumption 5.1.

CLT and Bootstrap CLT for $(\widehat\xi^d_{\bar y}, \widehat\xi^d_{\underline{y}}, \widehat\xi^d_{\bar w}, \widehat\xi^d_{\underline{w}})$. We use the notation R to refer to either Y and W, as the proof is identical in both cases. In what follows it is convenient to analyze the estimator of the lower tail, $\underline{r}$, the analysis for estimator of upper tail follows exactly the same steps, switching the sign of the dependent variable, $R \in \{Y, W, -Y, -W\}$.

The estimator $\widehat\xi_{\underline{r}}$ can be seen as a Z-estimator with moment function

$$\varphi_i(\alpha_{\underline{r}}, \mu_{\underline{r}}) = (\partial_\alpha \ell_i^{r_0}(\alpha_{\underline{r}}, \mu_{\underline{r}}), \partial_\mu \ell_i^{r}(\mu_{\underline{r}})')'.$$

Invoking Z-process theory for the simple case of finite-dimensional space $\mathbb{R}^{d_x+1}$, we have that jointly in $R \in \{Y, -Y, W, -W\}$,

$$\sqrt{n/n_d}\sqrt{n_d}\left(\widehat\xi^d_{\underline{r}} - \xi^d_{\underline{r}}\right) \rightsquigarrow \sqrt{p_d}\mathrm{E}\begin{bmatrix} \partial_{\alpha\alpha}\ell_i^{r_0}(\alpha^d_{\underline{r}}, \mu^d_{\underline{r}}) & \partial_{\alpha\mu}\ell_i^{r_0}(\alpha^d_{\underline{r}}, \mu^d_{\underline{r}}) \\ 0 & \partial_{\mu\mu}\ell_i^{r}(\mu^d_{\underline{r}}) \end{bmatrix}^{-1} \mathbb{G}(\varphi_i(\alpha^d_{\underline{r}}, \mu^d_{\underline{r}})). \quad \text{(A.3)}$$

In fact using Hadamard differentiability results for Z-processes given in Chernozhukov et al. (2013), we conclude that convergence results (A.2) and (A.3) hold jointly.[8]

The Hadamard differentiability results for Z-processes also imply that the bootstrap analogs of the results (A.2) and (A.3) are valid and hold jointly as well. We omit writing the formulas for these convergence results, since they are analogous to (A.2) and (A.3). ■

A.3. Proof of Theorem 5.1. Let $UC(\bar{\mathcal{Y}}\bar{\mathcal{W}}, \rho)$ denote the set of functions mapping $\bar{\mathcal{Y}}\bar{\mathcal{W}}$ to the real line that are uniformly continuous with respect to ρ, the standard metric in $\mathbb{R}^2$. The proof follows similar steps to the proof of Theorem 5.2 in Chernozhukov et al. (2013), suitably modified to extend the process to the tails. The main differences are highlighted in Steps 1, 2, and 3 below.

Step 1.(Results for coefficients and empirical measures). Application of the Hadamard differentiability results for Z-processes in Chernozhukov et al. (2013), together with

[8] Of course, it is cumbersome to put this joint convergence statement into one display, so we state this verbally.

Lemma 5.1, gives that,

$$\sqrt{n/n_m}\sqrt{n_m}\,\widehat{G}_X^m(f) \rightsquigarrow G_X^m(f) \text{ in } \ell^\infty(\mathcal{F}), \tag{A.4}$$

jointly with (A.2) and (A.3) for all $d, m \in \{0, 1\}$.

Step 2.(Main: Results for conditional cdfs). Here we show that, for all $j, k, l, m \in \{0, 1\}$,

$$(\widehat{Z}_{ywx}^{(jkl)}, \widehat{G}_X^m(f)) \rightsquigarrow (Z_{ywx}^{(jkl)}, G_X^m(f)) \text{ in } \ell^\infty(\mathcal{YWXF}),$$
$$(\widehat{Z}_{ywx}^{(jkl)*}, \widehat{G}_X^{m*}(f)) \rightsquigarrow_{\mathrm{P}} (Z_{ywx}^{(jkl)}, G_X^m(f)) \text{ in } \ell^\infty(\mathcal{YWXF}).$$

As in the proof of Lemma 5.1, we drop the group indexes i, j and k to lighten the notation, whenever it does not cause confusion.

It is convenient to extend the processes $\widehat{\mu}_y$, $\widehat{\nu}_w$ and $\widehat{\delta}_{yw}$ to the tails as

$$\widehat{\mu}_y = \widehat{\mu}_{\bar{y}_y} + (y - \bar{y}_y)\widehat{\alpha}_{\bar{y}_y} e_1,\ \widehat{\nu}_w = \widehat{\nu}_{\bar{w}_w} + (w - \bar{w}_w)\widehat{\alpha}_{\bar{w}_w} e_1,\ \widehat{\delta}_{yw} = \widehat{\delta}_{\bar{y}_y \bar{w}_w}, \quad y \in \mathcal{Y} \setminus \bar{\mathcal{Y}}, w \in \mathcal{W} \setminus \bar{\mathcal{W}},$$

where e_1 is a unitary d_x-vector with a one in the first component. Likewise, the coefficient functions are given by

$$\mu_y = \mu_{\bar{y}_y} + (y - \bar{y}_y)\alpha_{\bar{y}_y} e_1,\ \nu_w = \nu_{\bar{w}_w} + (w - \bar{w}_w)\alpha_{\bar{w}_w} e_1,\ \delta_{yw} = \delta_{\bar{y}_y \bar{w}_w}, \quad y \in \mathcal{Y} \setminus \bar{\mathcal{Y}}, w \in \mathcal{W} \setminus \bar{\mathcal{W}},$$

by assumption.

For the body part, $(y, w) \in \bar{\mathcal{Y}} \times \bar{\mathcal{W}}$, consider the mapping $\phi : \mathbb{D}_\phi \subset \ell^\infty(\bar{\mathcal{Y}}\bar{\mathcal{W}})^{d_\theta} \to \ell^\infty(\bar{\mathcal{Y}}\bar{\mathcal{W}}\mathcal{X})$, defined as

$$t \mapsto \phi(t), \quad \phi(t)(y, w, x) := \Phi_2\left(x' m_y, x' n_w, g(x' d_{yw})\right), \quad t_{yw} := (m_y, n_w, d_{yw}).$$

It is straighforward to deduce that this map is Hadamard differentiable at $t_{yw} = \theta_{yw}$ tangentially to $UC(\bar{\mathcal{Y}}\bar{\mathcal{W}}, \rho)^{d_\theta}$ with the derivative map given by:

$$v \mapsto \phi'(v), \quad \phi'(v)(y, w, x) = \partial_\theta \Phi_2(x'\mu_y, x'\nu_w; g(x'\delta_{yw})) \cdot v_{yw}.$$

The tail part can be divided in 3 regions: (I) $(\mathcal{Y} \setminus \bar{\mathcal{Y}}) \times (\mathcal{W} \setminus \bar{\mathcal{W}})$, (II) $\bar{\mathcal{Y}} \times (\mathcal{W} \setminus \bar{\mathcal{W}})$, and (III) $(\mathcal{Y} \setminus \bar{\mathcal{Y}}) \times \bar{\mathcal{W}}$. For $(y, w) \in I$, we consider the mapping $\phi_I : \mathbb{R}^{3d_x+2} \to \ell^\infty((\mathcal{Y} \setminus \bar{\mathcal{Y}})(\mathcal{W} \setminus \bar{\mathcal{W}})\mathcal{X})$ defined by:

$$(t, a, b) \mapsto \phi_I(t, a, b), \quad \phi_I(t, a, b)(y, w, x) := \Phi_2(x'm + (y - \bar{y}_y)a, x'n + (w - \bar{w}_w)b, g(x'd)),$$
$$t := (m, n, d).$$

This map is also Hadamard differentiable at $(t, a, b) = (\theta_{\bar{y}_y, \bar{w}_w}, \alpha_{\bar{y}_y}, \alpha_{\bar{w}_w})$ tangentially to the entire domain with the derivative $(v, h, u) \mapsto \phi_I'(v, h, u)$, where

$$
\begin{aligned}
\phi_I'(v,h,u)(y,w,x) &= \partial_\theta \Phi_2(x'\mu_{\bar{y}_y} + (y-\bar{y}_y)\alpha_{\bar{y}_y}, x'\nu_{\bar{w}_w} + (w-\bar{w}_w)\alpha_{\bar{w}_w}, g(x'\delta_{\bar{y}_y\bar{w}_w}))v_{\bar{y}_y\bar{w}_w} \\
&+ \partial_{\alpha_{\bar{y}_y}} \Phi_2(x'\mu_{\bar{y}_y} + (y-\bar{y}_y)\alpha_{\bar{y}_y}, x'\nu_{\bar{w}_w} + (w-\bar{w}_w)\alpha_{\bar{w}_w}, g(x'\delta_{\bar{y}_y\bar{w}_w})))(y-\bar{y}_y)h \\
&+ \partial_{\alpha_{\bar{w}_w}} \Phi_2(x'\mu_{\bar{y}_y} + (y-\bar{y}_y)\alpha_{\bar{y}_y}, x'\nu_{\bar{w}_w} + (w-\bar{w}_w)\alpha_{\bar{w}_w}, g(x'\delta_{\bar{y}_y\bar{w}_w})))(w-\bar{w}_w)u.
\end{aligned}
$$

The derivative is a bounded (continuous) linear operator (note that as $y \to \pm\infty$ or $w \to \pm\infty$, the derivative vanishes, with the linear growth factors $y - \bar{y}_w$ or $w - \bar{w}_w$ being dominated by the term $\partial_{\alpha_{\bar{y}_y}} \Phi_2$ or $\partial_{\alpha_{\bar{w}_w}} \Phi_2$ with exponential tails). Similarly, for $(y,w) \in II$, we consider the mapping $\phi_{II} : \ell^\infty(\bar{\mathcal{Y}})^{2d_x} \times \mathbb{R}^{d_x+1} \to \ell^\infty(\bar{\mathcal{Y}}(\mathcal{W} \setminus \bar{\mathcal{W}})\mathcal{X})$ defined by:

$$
(t,b) \mapsto \phi_{II}(t,b), \quad \phi_{II}(t,b)(y,w,x) := \Phi_2(x'm_y, x'n + (w-\bar{w}_w)b, g(x'd_y)),\ t_y := (m_y, n, d_y).
$$

This map is Hadamard differentiable at $(t_y, b) = (\theta_{y,\bar{w}_w}, \alpha_{\bar{w}_w})$ tangentially to the entire domain with the derivative $(v,u) \mapsto \phi_{II}'(v,u)$, where

$$
\begin{aligned}
\phi_{II}'(v,u)(y,w,x) &= \partial_\theta \Phi_2(x'\mu_y, x'\nu_{\bar{w}_w} + (w-\bar{w}_w)\alpha_{\bar{w}_w}, g(x'\delta_{y\bar{w}_w}))v_{y\bar{w}_w} \\
&+ \partial_{\alpha_{\bar{w}_w}} \Phi_2(x'\mu_y, x'\nu_{\bar{w}_w} + (w-\bar{w}_w)\alpha_{\bar{w}_w}, g(x'\delta_{y\bar{w}_w})))(w-\bar{w}_w)u.
\end{aligned}
$$

The derivative is a bounded (continuous) linear operator. Finally, for $(y,w) \in III$, we consider the mapping $\phi_{III} : \ell^\infty(\bar{\mathcal{W}})^{2d_x} \times \mathbb{R}^{d_x+1} \to \ell^\infty((\mathcal{Y} \setminus \bar{\mathcal{Y}})\bar{\mathcal{W}}\mathcal{X})$ defined by:

$$
(t,a) \mapsto \phi_{III}(t,a), \quad \phi_{III}(t,a)(y,w,x) := \Phi_2(x'm + (y-\bar{y}_y)a, x'n_w, g(x'd_w)),\ t_w := (m, n_w, d_w).
$$

This map is Hadamard differentiable at $(t_w, a) = (\theta_{\bar{y}_y,w}, \alpha_{\bar{y}_y})$ tangentially to the entire domain with the derivative $(v,h) \mapsto \phi_{III}'(v,h)$, where

$$
\begin{aligned}
\phi_{III}'(v,h)(y,w,x) &= \partial_\theta \Phi_2(x'\mu_{\bar{y}_y} + (y-\bar{y}_y)\alpha_{\bar{y}_y}, x'\nu_w, g(x'\delta_{\bar{y}_y w}))v_{\bar{y}_y w} \\
&+ \partial_{\alpha_{\bar{y}_y}} \Phi_2(x'\mu_{\bar{y}_y} + (y-\bar{y}_y)\alpha_{\bar{y}_y}, x'\nu_w, g(x'\delta_{\bar{y}_y w}))(y-\bar{y}_y)h.
\end{aligned}
$$

The derivative is a bounded (continuous) linear operator.

We can now define the "extended map" that combines the body and tail pieces $(t,a,b) \mapsto \bar{\phi}(t,a,b)$, where

$$
\begin{aligned}
\bar{\phi}(t,a,b)(y,w,x) &:= \phi(t)(y,w,x)1((y,w) \in \bar{\mathcal{Y}} \times \bar{\mathcal{W}}) + \phi_I(t,a,b)(y,w,x)1((y,w) \in I) \\
&+ \phi_{II}(t,b)(y,w,x)1((y,w) \in II) + \phi_{III}(t,a)(y,w,x)1((y,w) \in III).
\end{aligned}
$$

By combining the four differentiability results above, we can deduce that this map is Hadamard differentiable with the derivative map $(v,h,u) \mapsto \bar{\phi}'(v,h,u)$, where

$$
\begin{aligned}
\bar{\phi}'(v,h,u)(y,w,x) &:= \phi'(v)(y,w,x)1((y,w) \in \bar{\mathcal{Y}} \times \bar{\mathcal{W}}) + \phi_I(v,u,h)(y,w,x)1((y,w) \in I) \\
&+ \phi_{II}(v,u)(y,w,x)1((y,w) \in II) + \phi_{III}(v,h)(y,w,x)1((y,w) \in III).
\end{aligned}
$$

Then, the claim for the conditional process $\widehat{Z}^{(ijk)}_{ywx}$ follows by (A.2), (A.3) and the functional delta method, and expression of the limit process can be obtained after some algebra.

STEP 3. (Auxiliary: Donskerness). One key ingredient for the result is to show that $\mathcal{F}$ is a Dudley-Koltchinskii-Pollard (DKP) class, namely it has bounded uniform covering entroy integral and obeys standard measurability condition (Dudley's image-admissible Suslin condition). We omit any discussion of measurability in this paper, but we note that it trivially holds. The proof in Chernozhukov et al. (2013) relies on compactness of the set $\mathcal{YWX}$ and applies immediately only to $\bar{\mathcal{Y}}\bar{\mathcal{W}}\mathcal{X}$. We extend the result to $\mathcal{YWX}$. Note that $\mathcal{F} = \{F_{Y,W|X}(y, w \mid \cdot) : y \in \mathcal{Y}, w \in \mathcal{W}\}$ is a uniformly bounded "parametric" family indexed by $y \in \mathcal{Y}$ and $w \in \mathcal{W}$ that obeys $|F_{Y,W|X}(y, w \mid \cdot) - F_{Y,W|X}(y', w' \mid \cdot)| \leqslant L(|y - y^{'}| + |w - w^{'}|)$, given the assumption that the density functions $f_{Y|X}$ and $f_{W|X}$ are uniformly bounded by some constant L. This is enough to bound the covering numbers for the index set $\bar{\mathcal{Y}} \times \bar{\mathcal{W}}$, but is not enough to bound the covering number over the unbounded set $\mathcal{Y} \times \mathcal{W}$.

Under our modelling hypotheses, there exists a small enough constant $C > 0$ such that

$$F_{Y,W|X}(y, w \mid \cdot) \leqslant \exp(Cy);\;\; y < 0; \qquad 1 - F_{Y,W|X}(y, w \mid \cdot) \leqslant \exp(-Cy);\;\; y > 0;$$

and

$$F_{Y,W|X}(y, w \mid \cdot) \leqslant \exp(Cw);\;\; w < 0; \qquad 1 - F_{Y,W|X}(y, w \mid \cdot) \leqslant \exp(-Cw);\;\; w > 0.$$

Let $R_j = -M(\varepsilon) - \varepsilon/(2L) + j(\varepsilon/L)$, with $j = 0, ..., J$, where $J = \lceil 2M(\varepsilon)L/\varepsilon\rceil + 1$, $M(\varepsilon) = \log(1/\varepsilon)/C$ and $\lceil\cdot\rceil$ is the ceiling function. Let $R_{-1} = -\infty$ and $R_{J+1} = +\infty$. The sets $B_{jk} = \{F_{Y,W|X}(y, w \mid X) : R_j \leqslant y \leqslant R_{j+1}, R_k \leqslant w \leqslant R_{k+1}\}$ for $j, k \in \{-1, ..., J\}$ have the L^2 diameter of at most $\sqrt{2}\varepsilon$ independently of the distribution of F_X:

- Indeed by the previous paragraph, if $j, k \in \{0, ..., J - 1\}$, then the diameter of the set $B_{j,k}$ is at most $\sqrt{2}L(\varepsilon/L) = \sqrt{2}\varepsilon$.
- For $j \in \{-1, J\}$ or $k \in \{-1, J\}$, then any pair of conditional cdfs in the same ball obey:

$$|F_{Y,W|X}(y, w \mid \cdot) - F_{Y,W|X}(y', w' \mid \cdot)| \leqslant \exp(-M(\varepsilon)C) = \exp(-[\log(1/\varepsilon)/C]C) \leqslant \epsilon.$$

The number of sets is at most $[2\log(1/\varepsilon)L/(C\varepsilon)+5]^2$. It follows that the uniform covering number of the function set $\mathcal{F} = \{F_{Y,W|X}(y, w \mid X) : y \in \mathcal{Y}, w \in \mathcal{W}\}$ is bounded by $(1 + (1/\varepsilon^2))^2$ up to a constant that does not depend on the distribution of X. The set $\mathcal{F}$ is therefore a DKP class. ∎

References

Abramitzky, Ran, Leah Boustan, Elisa Jácome, and Santiago Pérez, "Intergenerational mobility of immigrants in the united states over two centuries," *American Economic Review*, 2021, *111* (2), 580–608.

Adermon, Adrian, Mikael Lindahl, and Daniel Waldenström, "Intergenerational wealth mobility and the role of inheritance: Evidence from multiple generations," *The Economic Journal*, 2018, *128* (612), F482–F513.

______, ______, **and Mårten Palme**, "Dynastic human capital, inequality, and intergenerational mobility," *American Economic Review*, 2021, *111* (5), 1523–48.

Bouzebda, Salim and Boutheina Nemouchi, "Central limit theorems for conditional empirical and conditional U-processes of stationary mixing sequences," *Mathematical Methods of Statistics*, 2019, *28*, 169–207.

Callaway, Brantly and Weige Huang, "Local intergenerational elasticities," *Economics Bulletin*, 2019, *39* (2), 919–928.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly, "Inference on counterfactual distributions," *Econometrica*, 2013, *81* (6), 2205–2268.

______, ______, **and Siyi Luo**, "Distribution regression with sample selection, with an application to wage decompositions in the UK," *arXiv preprint arXiv:1811.11603*, 2018.

______, ______, **Whitney Newey, Sami Stouli, and Francis Vella**, "Semiparametric estimation of structural functions in nonseparable triangular models," *Quantitative Economics*, 2020, *11* (2), 503–533.

______, **Iván Fernández-Val, Blaise Melly, and Kaspar Wüthrich**, "Generic inference on quantile and quantile effect functions for discrete outcomes," *Journal of the American Statistical Association*, 2020, *115* (529), 123–137.

______, ______, **Jonas Meier, Aico van Vuuren, and Francis Vella**, "Conditional rank-rank regression," 2025.

Chetty, Raj, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter, "Race and economic opportunity in the United States: An intergenerational perspective," *The Quarterly Journal of Economics*, 2020, *135* (2), 711–783.

______, ______, **Patrick Kline, and Emmanuel Saez**, "Where is the land of opportunity? The geography of intergenerational mobility in the United States," *The Quarterly Journal of Economics*, 2014, *129* (4), 1553–1623.

Fernandez-Val, Ivan, Wayne Yuan Gao, Yuan Liao, and Francis Vella, "Dynamic heterogeneous distribution regression panel models, with an application to labor income processes," 2023.

Fernández-Val, Ivan, Aico van Vuuren, and Francis Vella, "Nonseparable sample selection models with censored selection rules," *Journal of Econometrics*, 2024, *240* (2), 105088.

Fernández-Val, Iván, Aico van Vuuren, and Francis Vella, "Marital sorting, household inequality and selection," 2023.

———, **Franco Peracchi, Aico van Vuuren, and Francis Vella**, "Changes in the joint distribution of earnings for U.S. married couples, 1967–2019," 2022.

———, **Jonas Meier, Aico van Vuuren, and Francis Vella**, "Distribution regression difference-in-differences," 2024.

Foresi, Silverio and Franco Peracchi, "The conditional distribution of excess returns: an empirical analysis," *Journal of the American Statistical Association*, 1995, *90* (430), 451–466.

Klein, Nadja, Torsten Hothorn, Luisa Barbanti, and Thomas Kneib, "Multivariate conditional transformation models," *Scandinavian Journal of Statistics*, 2022, *49* (1), 116–142.

Landersø, Rasmus and James J Heckman, "The Scandinavian fantasy: the sources of intergenerational mobility in Denmark and the US," *The Scandinavian Journal of Economics*, 2017, *119* (1), 178–230.

Meier, Jonas, "Multivariate distribution regression," Technical Report, Universitaet Bern, Departement Volkswirtschaft 2020.

Mogstad, Magne and Gaute Torsvik, "Chapter 6 - Family background, neighborhoods, and intergenerational mobility," in "Handbook of the Economics of the Family, Volume 1," Vol. 1 of *Handbook of the Economics of the Family*, North-Holland, 2023, pp. 327–387.

van der Vaart, Aad W., *Asymptotic statistics* Cambridge Books, first edition ed., Cambridge University Press, September 2000.

van der Vaart, Aad W. and Jon Wellner, *Weak convergence and empirical processes: with applications to statistics* Springer Series in Statistics, first edition ed., Springer, 1996.

Wang, Yunyun, Tatsushi Oka, and Dan Zhu, "Bivariate distribution regression with application to insurance data," *arXiv preprint arXiv:2203.12228*, 2022.

Williams, O. Dale and James E. Grizzle, "Analysis of contingency tables having ordered response categories," *Journal of the American Statistical Association*, 1972, *67* (337), 55–63.

Appendix B. Robustness checks

FIGURE 4. Counterfactual transition matrices using no covariates for ρ (boostrapped standard errors in parentheses).

(A) Daughters

(B) Sons

(C) Daughters, $\rho = 0$

(D) Sons, $\rho = 0$

(E) Daughters, difference due to ρ

(F) Sons, difference due to ρ

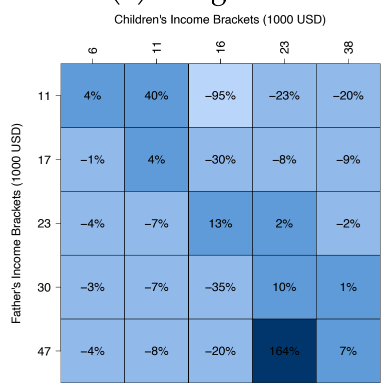
Notes: Panel (A) and (B) show the predicted transition matrices from the full model. For panels (C) and (D), $\widehat{\rho}$ is artificially set to zero. The transition matrices capture the dependence due to the marginals. Panel (E) and (F) show the differences between (A) and (C) and (B) and (D). Standard errors calculated as in Figure 1.

FIGURE 5. Decomposition of difference in transition matrices between sons and daughters excluding race and household size for children.

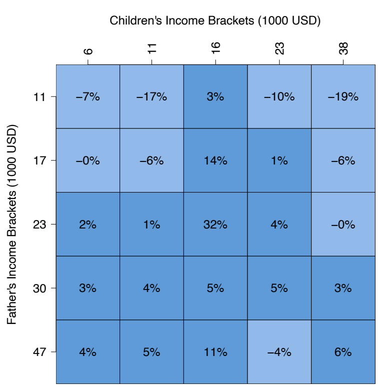
(A) Sons

(B) Daughters

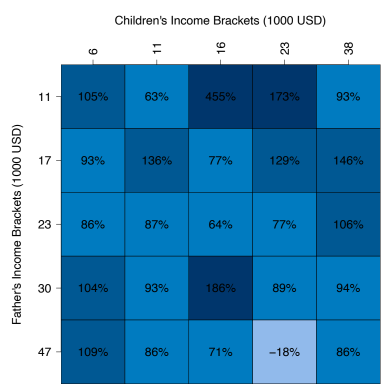
(C) Total Difference



(D) Composition



(E) Sorting



(F) Marginals

Notes: Figure 5 shows the decomposition of the joint distribution between sons and daughters into three terms.