Nonparametric Estimation of a Generalized Regression Model with Group Effects

EDUARDO A. SOUZA-RODRIGUES[*]
University of Toronto, Department of Economics

March 31, 2015

Abstract

This paper develops a nonparametric estimator for the generalized regression model proposed by Berry and Haile (2009) in which each individual is associated with a group and each group is subject to observable and unobservable shocks. The motivation for this model is to estimate the effects of group-level observables on individual outcomes when group-level observables correlate with group-level unobservables. Examples of groups include markets, regions and time periods, and group-level observables may include prices and policies. Group-level unobservables may be indexed by individual characteristics, which allows for more general group shocks than existing approaches. We propose a two-step estimator in which the first step runs a nonparametric regression of individual outcomes on individual observables within each group. It is a nonparametric regression in the presence of common shocks. The second step fixes the individual characteristics and runs a nonparametric quantile instrumental variable regression across groups of the predicted outcome obtained in the first step on group-level variables. It separates the effects of group-level observables from unobservables. We establish consistency and convergence rate of the estimator as well as the rates at which both the number of groups and the number of observations within each group have to increase to guarantee consistency.

JEL Classifications: C13, C14, C21.

KEYWORDS: Nonparametric regression, kernel regression, common shocks, penalized sieve minimum distance, quantile IV.

[*] University of Toronto, Department of Economics. E-mail: edusouzarod@gmail.com. Acknowledgements: I am deeply indebted to my committee members—Philip Haile, Steven Berry, Donald Andrews and Xiaohong Chen— for their continual support and encouragement. I also am very thankful to Yuichi Kitamura, Tai Otsu, Peter Phillips, Oliver Linton, Joseph Altonji, Ed Vytlacil and participants of Yale Econometrics Workshop. Finally, I thank Marfisa Queiroz for her support. Financial support from Charles V. Hickox Fellowship, Arthur K. Watson Fund, Yale University and Kernan Brothers Environmental are gratefully acknowledged. All errors are my responsibility.

1 Introduction

This paper develops a nonparametric estimator for a generalized regression model in which each individual is associated with a group and each group is subject to observable and unobservable shocks. The objective is to estimate the effects of group-level observables on individual outcomes when group-level observables correlate with group-level unobservables. The framework emcompasses many nonlinear models of interest such as binary choice and threshold crossing models, censored regression, and the proportional hazard model. The model can be applied to cross-sectional data with individuals observed in different groups, e.g., markets or regions; to repeated cross-sections; and to panel data. Group-level observables may include prices and policy interventions.

In the generalized regression model of Han (1987), an outcome Y_i of individual i is determined by the nonseparable model

$$\begin{aligned} Y_i &= D\left(Y_i^*\right) \\ Y_i^* &= G\left(S_i'\beta, \varepsilon_i\right), \end{aligned} \tag{1}$$

where $D(\cdot)$ is a known weakly increasing function; Y_i^* is the latent variable that depends on the observable covariates, S_i, and on the unobservable individual heterogeneity, ε_i; and β is the parameter of interest. Berry and Haile (2009) extend Han's model in several dimensions for panel settings in which each individual i is associated with a group t:[1]

$$\begin{aligned} Y_{it} &= D\left(Y_{it}^*\right) \\ Y_{it}^* &= G\left(S_{it}, X_t, U_t\left(S_{it}\right), \varepsilon_{it}\right), \end{aligned} \tag{2}$$

where X_t is the observable group-specific covariates, and $U_t\left(S_{it}\right)$ is the unobservable "group effect." Note first that the covariates do not enter through an index in $G\left(.\right)$. Instead, the model is fully nonparametric. Second, the model allows for heterogeneous responses across individuals to covariates. I.e., X_t can affect the entire distribution of Y_{it}^*, holding S_{it} and $U_t\left(S_{it}\right)$ fixed. This is not possible in models with a single stochastic element. In particular, (2) allows for random coefficient models. Finally, the unobservable group-effects, $U_{it}\left(S_{it}\right)$, can be correlated with X_t and is indexed by individual characteristics, S_{it}. This allows for more general group fixed-effects than existing approaches. For example, in repeated cross-sections, X_t may be a policy instrument common to all individuals in each time period and $U_t\left(S_{it}\right)$ may be an unobserved macroeconomic shock correlated with X_t and with impacts that vary continuously with, say, individuals' wealth, S_{it}.

[1] The usual indices for panel data are reversed here: for an individual observed in several time periods, we denote t for the individual and i for the different time periods. Each individual plays the role of a "group."

We propose a nonparametric estimator for two parameters of interest: (a) the conditional distribution of the observed outcome, $\Pr(Y_{it}|S_{it}, X_t, U_t(S_{it}))$; and (b) the conditional distribution of the latent outcome, $\Pr(Y_{it}^{*}|S_{it}, X_t, U_t(S_{it}))$. Formally, let the support of $(S_{it}, X_t, U_t(S_{it}), \varepsilon_{it})$ be denoted by the product $\mathcal{S} \times \mathcal{X} \times \mathcal{U} \times \mathcal{E}$ $(\subseteq \mathbb{R}^{d_S} \times \mathbb{R}^{d_X} \times \mathbb{R} \times \mathbb{R}^{d_\varepsilon})$, the support of Y_{it} by $\mathcal{Y}$ $(\subseteq \mathbb{R})$ and the support of Y_{it}^{*} by $\mathcal{Y}^{*}$ $(\subseteq \mathbb{R})$. To handle the endogeneity of X_t, we make use of instrumental variables $Z_t \in \mathcal{Z}$ $(\subseteq \mathbb{R}^{d_Z}$, with $d_Z \geq d_X)$. The data set is given by $\{(Y_{it}, S_{it}, X_t, Z_t) : i = 1, ..., N_t, t = 1, ..., T\}$.

To make the model more concrete, we provide below two detailed examples that motivate the present paper.[2]

Example 1 ***(Binary Choice Demand)****. Suppose a consumer i in market t decides whether to buy a product. Let v_1 denote the indirect utility from consuming the product and v_0 the utility of the outside option. The indirect utility may depend on consumer's observable characteristics, S_{it}, such as income and gender, and on unobservable heterogeneous tastes ε_{it}. It may also depend on observable characteristics of the product, X_t, such as price, and on the unobserved product quality, which may be correlated with price. Let the unobserved product quality be denoted by $U_t(S_{it})$ indicating that the product may have different appeal to consumers with different characteristics S_{it}. For example, women may value the unobservable quality, but men may not appreciate it as much. The latent utility difference is*

$$Y_{it}^{*} = v_1(S_{it}, X_t, U_t(S_{it}), \varepsilon_{it}) - v_0(S_{it}, \varepsilon_{it}), \tag{3}$$

and individual's zero-one purchase decision is given by $D(Y_{it}^{*}) = 1[Y_{it}^{*} \geq 0]$. A random coefficient model is obtained by imposing

$$Y_{it}^{*} = S_{it}'\gamma + X_t'\beta_{it} + U_t(S_{it}) + \eta_{it},$$

where $\varepsilon_{it} = (\beta_{it}, \eta_{it})$ can have arbitrary joint distribution and depend on S_{it}. The object of interest may be either the demand function $\Pr(Y_{it} = 1|S_{it}, X_t, U_t(S_{it}))$, or its price-elasticity, or the conditional distribution of utility differences, $\Pr(Y_{it}^{*} \leq y \mid S_{it}, X_t, U_t(S_{it}))$.

Example 2 ***(Patient Outcomes).*** *A problem of public interest is the relationship between hospital volumes of surgical procedures and individual mortality rates. Numerous studies have documented*

[2] Other possible applications include: (a) impacts of policies on crime [Durlauf, Navarro and Rivers (2010)]; (b) impacts of roads on individual land use decisions, in particular on deforestation [Souza-Rodrigues, (2014b)]; (c) effects of local expenditures on advertisement on voter's behavior; (d) how state taxes affect firms' investments; (e) firm's entry decision or technology adoption with market-specific unobservables; and (f) schooling decision with school-specific unobservables.

an inverse relationship, but the evidence is weak for most operations; see, e.g., Birkmeyer et al. (2002), Finks, Osborne and Birkmeyer (2011), and the literature cited therein. The studies that find an inverse relationship suggest that thousands of death per year in the United States could have been prevented if hospitals with inadequate experience (i.e., with low volume of operations) have performed less surgical procedures. To study such a relationship, consider a model of health outcomes of patients i treated in hospital t. The latent Y_{it}^* may represent a continuous measure of actual health status, and the observable outcome Y_{it} may be the binary indicator for death vs. survival after a surgical intervention. The vector S_{it} represents patient demographics, such as gender and age. The vector X_t includes hospital characteristic, such as the volume of operations, staffing ratios, size, and for-profit status. The unobserved $U_t(.)$ may reflect unmeasured hospital quality (resulting from, say, unmeasured quality of staff, equipements, etc.). It is indexed by S_{it} because an unobserved hospital characteristic that is helpful for patients with one S_{it} may be harmful to other patients. Finally, we need instrumental variables for hospital volumes because hospitals with high volumes very likely are those with high unmeasured quality. A potential candidate for an intrument is the number of hospital per-capita in the region. The number of hospital per-capita should affect the level of local competition, and so affect the equilibrium volumes of surgeries, but should be excluded from the individual mortality equation. To the best of our knowledge, the existing literature has not implemented any instrumental variable approach.

Berry and Haile (2009) provide conditions to identify both parameters $\Pr(Y_{it}|S_{it}, X_t, U_t(S_{it}))$ and $\Pr(Y_{it}^*|S_{it}, X_t, U_t(S_{it}))$; in the present paper we develop a nonparametric estimator based on Berry and Haile's (2009) insights. We propose a two-step estimator that exploits the presence of a group-level special regressor. The special regressor can be used to trace the conditional distribution of the individual latent outcome. As commonly assumed in the literature, it must satisfy a usual large support assumption and enter additively in the $G(.)$ function in (2). However, it does not have to be exogenous. More precisely, although it must be independent of ε_{it}, it does not have to be independent of $U_t(s)$, for any $S_{it} = s$.

The first step runs nonparametric regressions of individual outcomes Y_{it} on individual observables S_{it} for each group t. It is a nonparametric regression in the presence of common shocks, where the common shocks include the random function $U_t(.)$. Souza-Rodrigues (2013) studies the properties of a kernel regression in the presence of common shocks that covers the present case and that was not previously considered in the literature. The presence of the potentially infinite-dimensional object $U_t(.)$ common to all individuals in a group t has to be handled with care because conditional densities given $U_t(.)$ may not exist in this context. We present sufficient conditions for existence of conditional densities and find that the restrictions are mild: $U_t(\cdot)$ must belong to a separable metric space and it must be sufficiently smooth if S_{it} is continuously distributed (twice differentiability suffices). Although kernel asymptotic results have to adapted here, the implementation does not differ from standard kernel regression.

The second step fixes the individual characteristics $S_{it} = s$, and runs a nonparametric quantile instrumental variable regression (NPQIV) across groups of the predicted outcome obtained in the first step on group-level observables, X_t. It separates the effects of X_t from $U_t(s)$. The second step modifies the penalized sieve minimum distance estimator (PSMD), developed by Chen and Pouzo (2009, 2012), to take into account the preliminary estimator from the first step. The main difficulty in the second step comes from the fact that the criterion function is not differentiable with respect to the parameter of interest. Chen and Pouzo (2012) exploit a Lipschitz condition in the criterion function to obtain consistency and rate of convergence. In the present case however the preliminary estimator breaks the Lipschitz condition and complicates the proofs of the asymptotic results. The difficulty holds independently on how the preliminary estimator is obtained. In spite of this difficulty, we establish consistency and convergence rate of the estimator. Consistency is obtained regardless of how fast the number of observations within each group, N_t, increases relative to the number of groups, T, provided both numbers go to infinity. In this sense, the estimator is a "large-N, large-T" estimator. The convergence rate on the other hand depends on how fast N_t increases compared to T. Perhaps not surprising, the faster the N_t increases relatively to T, the faster the convergence rate.

This paper relates to a large literature including the *generalized regression models* (Han (1987), Manski (1985), Abrevaya (2000), Honore and Lewbel (2002), Ichimura and Thompson (1998), Chiappori, Komunjer and Kristensen (2011)); the *non-linear panel data* (Altonji and Matzkin (2005), Hoderlein and White (2010) and Evdokimov (2010)); the *nonseparable models with endogeneity* (Chernozhukov and Hansen (2005), Chernozhukov, Imbens and Newey (2007), Chen, Chernozhukov, Lee and Newey (2011), Torgovitsky (2010)); the *multinomial choice model* (Berry and Haile (2010a,b), Fox and Gandhi (2011)); the *large-N/large-T panel data* (Phillips and Moon (1999)); the *cross-section with common shocks* (Andrews (2005)); and the *sieves estimator* (Ai and Chen (2003), Chen and Pouzo (2009, 2012), Horowitz and Lee (2007)).

The paper is organized as follows. Section 2 discusses potential advantages in allowing the group-level unobservables to be indexed by S_{it}. Section 3 presents Berry and Haile's (2009) identification results. Section 4 exposits the estimator. In Subsection 4.1, we present results for the first step and in Subsection 4.2, for the second step. Section 5 reports a Monte Carlo exercise, and Section 6 presents an application to hospital data, in which we estimate the impacts of the number of surgical procedures in a hospital on individual survival rates. Section 7 concludes. An Appendix provides proofs of results stated in the paper and is divided as follows. First, it presents the proofs of the results for the first and second steps (Appendix A.1 and A.2, respectively). Then, it describes the probabilistic framework that justifies the approach taken in this paper (Appendix A.3).

Notation The support of X is $\mathcal{X} = \mathcal{X}^1 \times ... \times \mathcal{X}^d$. Let f_X denote the density of X, and $f_{Y/X}$ denote the conditional density of Y given X. Denote $L^p(\mathcal{X}, \mu)$ as the space of functions h such that $\left(\int |h(x)|^p d\mu(x)\right)^{1/p} < \infty$, where μ is a sigma-finite measure. Define the $L^p(\mathcal{X}, \mu)$-norm as $\|h\|^p_{L^p(\mathcal{X},\mu)} = \int |h(x)|^p d\mu(x)$; and the L^p-empirical norm as $\|h\|^p_n = \frac{1}{n}\sum_{i=1}^{n} |h(X_i)|^p$. Define also the sup-norm $\|h\|_\infty = \sup_x |h(x)|$, and the Euclidean norm, $\|x\|_E = \left(\sum_{i=1}^{d} x^2\right)^{1/2}$. Given a d-tuple $\delta = (\delta_1, ..., \delta_d)$ of nonnegative integers, let $[\delta] = \delta_1 + ... + \delta_d$ and let D^δ denote the differential operator $D^\delta = \frac{\partial^{[\delta]}}{\partial x_1^{\delta_1} ... \partial x_d^{\delta_d}}$. For any positive sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \asymp b_n$ means that there exists positive constants c_1, c_2 such that $c_1 a_n \leq b_n \leq c_2 a_n$. The expression $a_n \lesssim b_n$ means there is a constant c such that $a_n \leq c b_n$. The term $a_n = O_p(b_n)$ means that for a positive constant M, $\lim_{M\to\infty} \limsup_{n\to\infty} \Pr(a_n/b_n > M) = 0$; and the term $a_n = o_p(b_n)$ means that for all $\varepsilon > 0$, we have $\lim_{n\to\infty} \Pr(a_n/b_n > \varepsilon) = 0$.

2 Group-Effect as a Random Function

The "group-effect", $U_t(S_{it})$, is allowed to be a random function of the individual covariates, S_{it}. This is more general than the common approach of allowing a unobserved group-level random variable, V_t, correlated with S_{it}. To discuss the potential gains of using a random function, we begin with an example. Suppose $S_{it} \in \{0, 1\}$. Then we can write

$$U_t(S_{it}) = U_{0t} \times 1\{S_{it} = 0\} + U_{1t} \times 1\{S_{it} = 1\}, \tag{4}$$

where $1\{.\}$ is an indicator function and the random variables U_{0t} and U_{1t} may or may not be independent to each other. We therefore split the original group t into two subgroups and allow for different "fixed-effects" affecting each. The observable X_t can be correlated with the vector (U_{0t}, U_{1t}) while Z_t must be independent of $(U_{0t}, U_{1t}, \varepsilon_{it})$. In Example 1, the binary choice demand example, (U_{0t}, U_{1t}) might capture how men and women rank differently the unoberevd quality of the good. The ranks could be completely different and independent to each other, a possibility that is not allowed if we restrict the group-level unobservables to the scalar V_t. There is no reason to believe a priori that these ranks should coincide.

In the binary S_{it} example, model (2) can be rewritten as

$$\begin{aligned} Y_{it}^{*} &= G\left(S_{it}, X_t, U_t\left(S_{it}\right), \varepsilon_{it}\right) \\ &= \widetilde{G}\left(S_{it}, X_t, U_{0t}, U_{1t}, \varepsilon_{it}\right). \end{aligned} \tag{5}$$

The literature on nonparametric identification, on the other hand, typically assumes a structure of the type

$$Y_{it}^{*} = \overline{G}\left(S_{it}, X_t, V_t, \varepsilon_{it}\right) \tag{6}$$

where $\overline{G}(\cdot)$ is typically strictly increasing in V_t; see, e.g., Altonji and Matzkin (2005), Hoderlein and White (2010) and Evdokimov (2010). The only way in which (5) can be reduced to (6) with $\overline{G}(\cdot)$ *strictly increasing* in V_t occurs when $U_{0t} = U_{1t} = V_t$. We do not rule out this possibility, but we allow for cases where $U_{0t} \neq U_{1t}$. Whether the general or the restrictive case is more reasonable depends on the application at hand.

It turns out that there exists a simple solution to handle the general case: allow $G(\cdot)$ to be strictly increasing in $U_t(S_{it})$ and carry out the entire analysis conditional on the event $\{S_{it} = s\}$. Conditional on $\{S_{it} = 0\}$, for example, equation (5) reduces to

$$\begin{aligned} Y_{it}^{*} &= G\left(0, X_t, U_t(0), \varepsilon_{it}\right) \\ &= G\left(0, X_t, U_{0t}, \varepsilon_{it}\right) \end{aligned} \tag{7}$$

with $G(\cdot)$ strictly increasing in $U_t(0)$. It is possible now to use standard identification results to recover the objects of interest. The price to be paid is the difficulty in estimating (and interpreting) the effects of S_{it} on outcomes, as will be clear in Section 3.[3]

The same reasoning applied to the binary S_{it} example can be extended to S_{it} taking finitely many or uncountably many values. The extension introduces some difficulties and requires extra notation because $U_t(\cdot)$ can be an infinite dimensional object. Formally, let $\mathcal{J}(\mathcal{S})$ be a space of

[3] Note that because we allow (U_{0t}, U_{1t}) to be dependent on each other, we can still accommodate a group-level fixed effect as usual. For example, instead of (4) we can assume

$$\begin{aligned} U_t(0) &= \alpha_0 V_t + U_{0t}, \\ U_t(1) &= \alpha_1 V_t + U_{1t}. \end{aligned} \tag{8}$$

where (U_{0t}, U_{1t}, V_t) may be jointly dependent or independent. Provided $G(\cdot)$ is strictly increasing in $U_t(S_{it})$, it is possible to let $\alpha_0 < 0 < \alpha_1$. In this case, the group-level fixed effect V_t may have negative impacts on individuals with demographics $S_{it} = 0$, while having positive impacts on individuals with $S_{it} = 1$.

functions mapping $\mathcal{S}$ into $\mathcal{U}$. We endow $\mathcal{J}(\mathcal{S})$ with an appropriate norm $\|\cdot\|_{\mathcal{J}}$ so that $\left(\mathcal{J}(\mathcal{S}), \|\cdot\|_{\mathcal{J}}\right)$ is a metric space and we equip this space with its Borel sigma-field, $\Im$. For each group t, we assume the function $U_t(\cdot)$ is a random object defined on the measurable space $(\mathcal{J}(\mathcal{S}), \Im)$.

Some restrictions on the space $\mathcal{J}(\mathcal{S})$ are necessary to obtain asymptotic results. As will be discussed in Section 4.1, the main restrictions we need are: (i) $\left(\mathcal{J}(\mathcal{S}), \|\cdot\|_{\mathcal{J}}\right)$ is a *separable* metric space equipped with its Borel σ-field, $\Im$, and (ii) the space $\mathcal{J}(\mathcal{S})$ is a subset of the twice continuously differentiable functions when S_{it} takes uncountable many values. These are mild restrictions.

Remark 1 *To illustrate the case where S_{it} is continuous, let $\mathcal{J}(\mathcal{S})$ be the Hilbert space $\mathcal{L}^2(\mathcal{S})$. Take a basis $\left\{\psi_j\right\}_{j=1}^{\infty}$ for $\mathcal{L}^2(\mathcal{S})$ and represent the group-effect by $U_t(S) = \sum_{j=1}^{\infty} V_{jt}\psi_j(S)$, where $V_{jt} \in \mathbb{R}$, for $j \geq 1$. The randomness of $U_t(S_{it})$ comes from the randomness of the infinite dimensional vector $(S_{it}, V_{1t}, V_{2t}, ...)$. We therefore can write model (2) as*

$$\begin{aligned} Y_{it}^* &= G\left(S_{it}, X_t, U_t\left(S_{it}\right), \varepsilon_{it}\right) \\ &= \widetilde{G}\left(S_{it}, X_t, \left(V_{1t}, V_{2t}, ...\right), \varepsilon_{it}\right). \end{aligned} \tag{9}$$

It is possible to represent the model using a single scalar V_t in place of the random vector $(V_{1t}, V_{2t}, ...)$, i.e., there exists V_t and $\overline{G}(\cdot)$ such that [Su, Hordelein and White (2010)]

$$\widetilde{G}\left(S_{it}, X_t, \left(V_{1t}, V_{2t}, ...\right), \varepsilon_{it}\right) = \overline{G}\left(S_{it}, X_t, V_t, \varepsilon_{it}\right).$$

However, there is no guarantee that $\overline{G}(\cdot)$ will be strictly increasing in V_t. The lack of monotonicity of $\overline{G}(\cdot)$ in V_t makes it difficult to identify the parameters of interest.

When S_{it} is continuous, the same simple solution to handle the binary S_{it} case can be applied: assume $G(\cdot)$ is strictly increasing in $U_t(S_{it})$ and carry out the analysis conditional on the event $\{S_{it} = s\}$.

Note that we cannot allow $U_t(S_{it})$ to be an unknown arbitrary deterministic function of S_{it}. In this case, model (2) reduces to

$$\begin{aligned} Y_{it}^* &= G\left(S_{it}, X_t, U_t\left(S_{it}\right), \varepsilon_{it}\right) \\ &= G_t\left(S_{it}, X_t, \varepsilon_{it}\right), \end{aligned} \tag{10}$$

and, because G_t could be arbitrarily different for different groups t, it would be impossible to identify effects of X_t on the outcomes.

Finally, an important observation that will be used in the rest of the paper is that all individuals in group t are affected by the common factors $(X_t, U_t(\cdot), Z_t) \in \mathcal{X} \times \mathcal{J}(\mathcal{S}) \times \mathcal{Z}$. We denote the common factors by $C_t = (X_t, U_t(\cdot), Z_t)$ and let $\mathcal{C} = \mathcal{X} \times \mathcal{J}(\mathcal{S}) \times \mathcal{Z}$. In case of a binary S_{it}, the common shock is the vector $C_t = (X_t, U_{0t}, U_{1t}, Z_t)$. In case S_{it} is continuous, the entire function $U_t(\cdot)$ is part of the common shock of group t, while $U_t(s)$ is a random variable that only affects those in group t with $S_{it} = s$. Note also that the group-level Z_t is part of the common shocks since it affects all individuals in group t, despite the fact it has indirect effects on Y_{it}^*.

3 Identification

The estimation procedure we propose follows closely the identification results of Berry and Haile (2009). They impose the following:[4]

Assumption 1 *For all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, there exists a known $\widetilde{y} \in \mathbb{R}$ such that,*

$$\Pr\left[D\left(G\left(S_{it}, X_t, U_t\left(S_{it}\right), \varepsilon_{it}\right)\right) \leq \widetilde{y} \mid S_{it} = s, X_t = x, U_t\left(s\right) = u\right] \textit{ is strictly decreasing in } u.$$

Sufficient for this condition is to assume that $G(\cdot)$ is strictly increasing in $U_t(\cdot)$. In the binary choice model, $Y_{it} = 1\left[Y_{it}^* \geq 0\right]$, we take $\widetilde{y} = 0$.

The second assumption states which distributions are known. We assume that (i) the joint distribution of $(Y_{it}, S_{it}, X_t, Z_t)$ is known and that (ii) the distribution of individual observable variables (Y_{it}, S_{it}) in a given group t is known. We denote the distribution within group t by $\Pr(\cdot|\ \mathrm{t})$.

Assumption 2 *(i) The joint distribution* $\Pr(Y_{it}, S_{it}, X_t, Z_t)$ *is known; and*

(ii) For all $y \in \mathcal{Y}$ and $s \in \mathcal{S}$, the distribution $\Pr(Y_{it} \leq y, S_{it} \leq s \mid \mathrm{t})$ *is known for each t.*

The common shock affecting group t is the random $C_t = (X_t, U_t(\cdot), Z_t) \in \mathcal{C}$. Define the measurable space $(\mathcal{C}, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-field. Let the distribution of C_t be denoted by Q_t (defined on $(\mathcal{C}, \mathcal{B})$) and define the σ-field generated by the common shock by $\sigma(C_t) = \sigma(X_t, U_t(\cdot), Z_t)$. Note that all that can be learned from the observations (Y_{it}, S_{it}) within the group t is the conditional distribution of (Y_{it}, S_{it}) given the sub-sigma-field $\sigma(C_t)$. It is not possible to recover the *unconditional* distribution of (Y_{it}, S_{it}) in group t except when (Y_{it}, S_{it}) is independent of C_t. But this independence is ruled out by assumption. Assuming the distribution of (Y_{it}, S_{it}) in group t is known is therefore equivalent to assuming that the conditional distribution (Y_{it}, S_{it}) given the

[4] For a complete discussion of the assumptions and the proof of identification the reader is referred to Berry and Haile (2009).

sub-sigma-field $\sigma(C_t)$ is known, even though the last object seems very abstract. In sum, we have that

$$\Pr(Y_{it} \leq y, S_{it} \leq s \mid \mathrm{t}) = \Pr(Y_{it} \leq y, S_{it} \leq s \mid \sigma(C_t)). \tag{11}$$

A formal discussion about the underlying probability space that justifies Assumption 2(ii) and the equality (11) is given in Appendix A.3 (see Lemma 3 in the Appendix A.3 and the discussion that follows it).

The next assumption requires fully independent instruments as usual in non-linear models:

Assumption 3 $(U_t(S_{it}), \varepsilon_{it}) \perp Z_t \mid S_{it}$.

Define

$$P_t(s, C_t) \equiv \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = s, \sigma(C_t)),$$

where $\widetilde{y}$ is the value referred in Assumption 1. By Assumption 2(ii), $P_t(s, C_t)$ is known. Moreover, by construction, we have that

$$\begin{aligned} P_t(s, C_t) &\equiv \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = s, X_t, U_t(\cdot), Z_t) \\ &= \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = s, X_t, U_t(s)) \\ &\equiv h_s(X_t, U_t(s)). \end{aligned} \tag{12}$$

where the first equality holds by the definition of $\sigma(C_t)$; the second equality is the result of two facts: (i) the variables Y_{it} and Z_t are independent given $(S_{it}, X_t, U_t(S_{it}))$; and (ii) conditioning on the event $\{S_{it} = s\} \cap \{X_t = x, U_t(\cdot) = u(\cdot)\}$ is equivalent to conditioning on $\{S_{it} = s\} \cap \{X_t = x, U_t(s) = u(s)\}$ (this is proved in Appendix A.3). The third line defines the function $h_s(\cdot)$. Note that, by Assumption 1, $h_s(x, u)$ is strictly decreasing in u.

Equation (12) is the fundamental equation of this paper. The identification and estimation results depend on this relationship. Because (i) $P_t(s, C_t)$ is known, (ii) $h_s(x, u)$ is strictly decreasing in u, and (iii) Z_t and $U_t(s)$ are independent random variables, the function $h_s(\cdot)$ can be nonparametrically identified using Chernozhukov and Hansen's (2005) identification results.

Formally, the unobservable $U_t(s)$ can be normalized to have uniform distribution over [0,1]. Once that h_s is identified, we can invert it to recover the *normalized* $U_t(s)$, i.e.,

$$U_t(s) = h_s^{-1}(X_t, P_t(s, C_t)),$$

where h_s^{-1} is the inverse of h_s. After recovering the normalized $U_t(s)$ for each s, we can identify the structural outcome distribution, $\Pr(Y_{it} | S_{it} = s, X_t = x, U_t(s) = u)$. That is precisely the approach adopted in Berry and Haile (2009).[5]

The next assumption is a "non-linear bounded completeness" assumption that is sufficient for Chernozhukov and Hansen's (2005) results. From now on, we normalize the distribution of $U_t(s)$ to be uniform on [0,1], for each $s \in \mathcal{S}$.

Assumption 4 *(i) Let $\mathcal{L}(u)$ be the convex hull of functions $m(x,u)$ satisfying:*

(a) *for all $z \in Z$, $\Pr\left(h_s\left(X_t, U_t\left(s\right)\right) \leq m\left(X_t, u\right) \mid Z_t = z\right) \in [u - \varepsilon_q, u + \varepsilon_q]$, for $\varepsilon_q > 0$; and*

(b) *for all $x \in X$, $m\left(x, u\right) \in p_x \equiv \left\{h_s : f_{P_s|X,Z}\left(h_s|x,z\right) \geq \varepsilon_f > 0, \forall z \text{ with } f_{X|Z}\left(x|z\right) > 0\right\}$, where $f_{P_s|X,Z}$ is the conditional density of $P_t\left(s\right)$ given $\left(X_t, Z_t\right)$ and $f_{X|Z}$ is the conditional density of X_t given Z_t.*

(ii) For all (x,u), assume $h_s(x,u) \in p_x$.

(iii) Define $\epsilon_t \equiv h_s(X_t, U_t\left(s\right)) - h_s(X_t, u)$ and let the density of ϵ_t be bounded and continuous on $\mathbb{R}$ a.s..

(iv) For any $u \in (0,1)$, for any bounded function $B(x,u) = m(x,u) - h_s(x,u)$ with $m(\cdot,u) \in \mathcal{L}(u)$, assume

$$E\left[B(X_t,u)\psi(X_t,Z_t,u)|Z_t\right] = 0 \ a.s.$$

only if $B(X_t,u) = 0$ a.s. for $\psi(x,z,u) = \int_0^1 f_{\epsilon_t}\left(\delta B(x,u)|x,z\right) d\delta$.

Based on this reasoning, Berry and Haile (2009) prove the following theorem:

Theorem 1 *Suppose Assumptions 1-4 hold, then the conditional probability* $\Pr\left(Y_{it}|S_{it}, X_t, U_t\left(S_{it}\right)\right)$ *is identified.*

To identify the conditional distribution $\Pr\left(Y^*_{it}|S_{it}, X_t, U_t\left(S_{it}\right)\right)$, they introduce a group-level special regressor. Formally, they partition X_t as $\left(X_t^{(1)}, X_t^{(2)}\right) \in \mathcal{X}^{(1)} \times \mathcal{X}^{(2)}$ and proceed with the following assumptions:

Assumption 5 *(Large Support) Supp$\left(X_t^{(1)}|S_{it}, X_t^{(2)}, U_t\left(S_{it}\right)\right) = \mathbb{R}$.*

Assumption 6 *(Separability) $G\left(S_{it}, X_t, U_t\left(S_{it}\right), \varepsilon_{it}\right) = X_t^{(1)} + g\left(S_{it}, X_t^{(2)}, U_t\left(S_{it}\right), \varepsilon_{it}\right)$.*

Assumption 7 *(Conditional Independence) $\varepsilon_{it} \perp X_t^{(1)} \mid S_{it}, X_t^{(2)}, U_t\left(S_{it}\right)$.*

[5] Although we could have different instruments $Z_t\left(s\right)$ corresponding to different demographics $S_{it} = s$, we do not exploit that possibility here.

Assumptions 5 and 6 are standard for models with a special regressor. Assumption 7 allows the special regressor to be correlated with group-effects, so $X_t^{(1)}$ does not have to be exogenous. The following theorem is then proved by Berry and Haile (2009):

Theorem 2 *Suppose Assumptions 1-7 hold, then the probability distribution* $\Pr\left(Y_{it}^{*}|S_{it},X_t,U_t\left(S_{it}\right)\right)$ *is identified.*

Next note a relationship between the distribution $\Pr\left(Y_{it}^{*}\leq y\mid S_{it}=s,X_t=x,U_t\left(s\right)=u\right)$ and the function $h_s\left(x,u\right)$ that simplifies the estimation procedure considerably. Define $D^{-1}\left(y\right)=\sup\left\{q:D\left(q\right)\leq y\right\}$ and let $\overline{x}^{(1)}$ be such that $\left(y-x^{(1)}\right)=\left(D^{-1}\left(\widetilde{y}\right)-\overline{x}^{(1)}\right)$, for a given $\left(y,x^{(1)}\right)$. For any pair $\left(y,x^{(1)}\right)$, a corresponding $\overline{x}^{(1)}$ exists by Assumption 5.

We have then the sequence of equalities:

$$
\begin{aligned}
& \Pr\left(Y_{it}^{*}\leq y\mid S_{it}=s,X_t^{(1)}=x^{(1)},X_t^{(2)}=x^{(2)},U_t\left(s\right)=u\right) \\
= \; & \Pr\left(X_t^{(1)}+g\left(S_{it},X_t^{(2)},U_t\left(S_{it}\right),\varepsilon_{it}\right)\leq y\mid S_{it}=s,X_t^{(1)}=x^{(1)},X_t^{(2)}=x^{(2)},U_t\left(s\right)=u\right) \\
= \; & \Pr\left(g\left(S_{it},X_t^{(2)},U_t\left(S_{it}\right),\varepsilon_{it}\right)\leq y-x_t^{(1)}\mid S_{it}=s,X_t^{(2)}=x^{(2)},U_t\left(s\right)=u\right) \\
= \; & \Pr\left(g\left(S_{it},X_t^{(2)},U_t\left(S_{it}\right),\varepsilon_{it}\right)\leq D^{-1}\left(\widetilde{y}\right)-\overline{x}^{(1)}\mid S_{it}=s,X_t^{(2)}=x^{(2)},U_t\left(s\right)=u\right) \\
= \; & \Pr\left(X_t^{(1)}+g\left(S_{it},X_t^{(2)},U_t\left(S_{it}\right),\varepsilon_{it}\right)\leq D^{-1}\left(\widetilde{y}\right)\mid S_{it}=s,X_t^{(1)}=\overline{x}^{(1)},X_t^{(2)}=x^{(2)},U_t\left(s\right)=u\right) \\
= \; & \Pr\left(Y_{it}\leq\widetilde{y}\mid S_{it}=s,X_t^{(1)}=\overline{x}^{(1)},X_t^{(2)}=x,U_t\left(s\right)=u\right) \\
= \; & h_s\left(\overline{x}^{(1)},x^{(2)},u\right) \qquad (13)
\end{aligned}
$$

where the first equality comes from Assumption 6 (separability); the second equality, from Assumption 7 (conditional independence); the third equality, from Assumption 5 (large support); the fifth equality, from the model (2); and the last equality, from the definition of h_s.

Hence the distribution function of Y_{it}^{*} evaluated at y and conditioned on $\left(s,x^{(1)},x^{(2)},u\right)$ equals the function h_s evaluated at the point $\left(\overline{x}^{(1)},x^{(2)},u\right)$, where $\overline{x}^{(1)}=x^{(1)}+D^{-1}\left(\widetilde{y}\right)-y$. All information about the distribution of the latent variable is therefore contained in the function h_s; we can estimate this conditional distribution directly using a quantile IV estimator $\widehat{h}_s$. Note that different than usual, the quantile approach here is used to identify a distribution function directly, and not its inverse. I.e., it identifies $\Pr\left(Y_{it}^{*} \leq y \mid S_{it}, X_t, U_t\left(S_{it}\right)\right)$, and not the quantile of Y_{it}^{*} given $\left(S_{it}, X_t, U_t\left(S_{it}\right)\right)$.

To identify the outcome distribution $\Pr\left(Y_{it} | S_{it}, X_t, U_t\left(S_{it}\right)\right)$, we can put (13) together with the definition $Y_{it} = D\left(Y_{it}^{*}\right)$ to obtain

$$\begin{aligned} & \Pr\left(Y_{it} \leq y \mid S_{it} = s, X_t = x, U_t\left(s\right) = u\right) \\ = \quad & \Pr\left(Y_{it}^{*} \leq D^{-1}\left(y\right) \mid S_{it} = s, X_t = x, U_t\left(s\right) = u\right) \\ = \quad & h_s\left(x^{(1)} + D^{-1}\left(\widetilde{y}\right) - D^{-1}\left(y\right), x^{(2)}, u\right). \end{aligned}$$

Finally, if the object of interest is the effect of the demographics S_{it} on Y_{it}^{*}, instead of the effect of X_t, then, the researcher can estimate the difference

$$h_{s'}\left(x, u\right) - h_s\left(x, u\right)$$

for $s \neq s'$. Some caution is needed in how to interpret this object, though. If the group-effect were a random variable, V_t, this difference would capture the effect of S_{it} keeping the group-effect constant. However, when the group-effect is a random function and $U_t\left(s\right)$ is normalized to have uniform distribution, then the equality $U_t\left(s\right) = U_t\left(s'\right) = u$ fixes the *quantiles* of the *non-normalized* $U_t\left(s\right)$ and $U_t\left(s'\right)$. I.e., by moving s to s', we are also moving the value of the non-normalized $U_t\left(s\right)$ from its u-quantile to the value of the same u-quantile of the non-normalized $U_t\left(s'\right)$. In this sense, all we can estimate is the across-groups quantile effect of S_{it}.

4 Estimator

The estimator relies on the equations (12) and (13). In the first step, we estimate

$$P_t\left(s, C_t\right) \equiv \Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it} = s, \sigma\left(C_t\right)\right),$$

for each group t. This requires a nonparametric regression in the presence of common shocks. An approach similar to that employed by Andrews (2005) for the linear regression model is adapted and applied here. Souza-Rodrigues (2013) studies the properties of the kernel regression estimator in the presence of common shocks that covers the present case when S_{it} is continuously distributed.

In the second step, instead of fixing the group t (or conditioning on $\sigma\left(C_t\right)$), we fix s and let $\left(X_t, U_t\left(s\right)\right)$ vary across t. It separates the effects of X_t and $U_t\left(s\right)$ on individual outcomes. We treat

$P_t(s, C_t) = h_s(X_t, U_t(s))$ as a nonparametric quantile IV model (NPQIV) and estimate h_s using the penalized sieve minimum distance estimator (PSMD) developed by Chen and Pouzo (2009, 2012). We have to modify the PSMD estimator to take into account the preliminary estimator $\widehat{P}_t(s)$ for $P_t(s, C_t)$. We show that the estimator is consistent provided both the number of groups, T, and the number of observations within groups, N_t, go to infinity. It is possible to obtain consistency when N_t is smaller than T, but the price to be paid is a slower convergence rate.[6]

4.1 First Step: Within Groups Estimator

Similar to Andrews (2005), we assume the cross-sectional dependence within group t only results from the common factors C_t.

Condition 1 *For each group t, $\{Y_{it}, S_{it} : i \geq 1\}$ are i.i.d. conditional on $\sigma(C_t)$.*

Condition 1 follows from Assumption A.1 and Corollary 1 in the Appendix A.3. Here we consider two cases: S_{it} is discrete and S_{it} is continuous.

4.1.1 Discrete Case

When S_{it} is discrete, the parametric approach employed by Andrews (2005) for the linear regression model is directly applicable. Suppose S_{it} can take finitely many values, say $S_{it} \in \{0, 1, .., L\}$. To make the connection with Andrews (2005) explicit, define, for $l = 1, .., L$,

$$S_{it}^l = \begin{cases} 0 \text{ if } S_{it} \neq l \\ 1 \text{ if } S_{it} = l \end{cases}$$

and define $\Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it}^1, ..., S_{it}^L, C_t\right)$ by

$$\Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it}^1, ..., S_{it}^L, C_t\right) = \beta_0(C_t) + \beta_1(C_t) \times S_{it}^1 + ... + \beta_L(C_t) \times S_{it}^L$$

where $\beta_l(C_t)$ are random coefficients measurable with respect to C_t. We have therefore

$$\begin{aligned} \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = 0, C_t) &= \beta_0(C_t) \\ \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = l, C_t) &= \beta_0(C_t) + \beta_l(C_t), \end{aligned}$$

[6] Note that it is not possible to estimate $\Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = s, X_t, U_t(s))$ directly using both cross-sectional and group variation simultaneously. The reason is that once we have cross-sectional observed variation S_{it}, we also should consider the presence of the unobserved heterogeneity ε_{it}. But once the individual heterogeneity is included, it would be impossible to separate it from the unobservable $U_t(\cdot)$ and, so, it would be impossible to invert an equation monotonic in $U_t(\cdot)$ and identify the "group-effect". On the other hand, if s is fixed and we explore the group-level variation, then all the unobservables across groups is captured by the scalar $U_t(s)$ and it is possible to identify it. That is the reason why we have to break the approach into two steps: the first one exploring within group variation and the second one exploring across group variation. In some situations, the researcher may not have access to the micro-data. In this case, it is still possible to use a restricted version of this estimator provided one observes the covariates X_t and some measure of $P_t(s)$. But the measure of $P_t(s)$ may still have effects on the rate of convergence and the asymptotic distribution.

for $l = 1, .., L$.

For each group t, we run a linear regression of $1\{Y_{it} \leq \widetilde{y}\}$ on $S_{it}^1, ..., S_{it}^L$ and take the estimated prediction of $1\{Y_{it} \leq \widetilde{y}\}$ given $S_{it} = s$ as our estimator $\widehat{P}_t(s)$, i.e., $\widehat{P}_t(s) = \widehat{\beta}_0 + \widehat{\beta}_s$.

Under some assumptions on existence of moments [see Assumptions 2 and 3 in Andrews (2005)], it is possible to show that $\widehat{\beta}_l \rightarrow_p \beta_l(C_t)$, for $l = 0, 1, ..., L$, and that the parametric rate of convergence is achievable [see Theorem 4 in Andrews (2005)]. More important for our objectives here is to show that

$$E\left(\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 \mid C_t\right) = e_t(s, C_t) \times N_t^{-1}$$

where N_t is the number of observations within group t and $e_t(s, C_t)$ is a $\sigma(C_t)$-measurable random variable that is almost surely finite. This result is obtained in Proposition 1 and is used in the second step of the estimator.

Proposition 1 *Let Condition 1 in the main text and Condition 9 in Appendix A.1 hold. Then,*

$$E\left(\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 \mid C_t\right) = e_t(s, C_t) \times N_t^{-1} \tag{14}$$

where $e_t(s, C_t)$ is a $\sigma(C_t)$-measurable random variable almost surely finite that is defined by (29) in Appendix A.1.

4.1.2 Continuous Case

When S_{it} is continuous, we run a nonparametric kernel regression of $1\{Y_{it} \leq \widetilde{y}\}$ on S_{it} for each group t. The Nadaraya-Watson kernel estimator is:

$$\widehat{P}_t(s) = \frac{\sum_{i=1}^{N_t} 1\{Y_{it} \leq \widetilde{y}\} K\left(\frac{S_{it}-s}{b}\right)}{\sum_{i=1}^{N_t} K\left(\frac{S_{it}-s}{b}\right)} \tag{15}$$

where $K(\cdot)$ is the kernel function and b is the bandwidth.

The literature on kernel estimators manipulates conditional densities of random variables, but it is well-known that conditional densities do not necessarily exist. In the present case the conditioning argument involves the infinite-dimensional function $U_t(.)$. If $U_t(.)$ is not restricted to a suitable space with a carefully constructed sigma-field, the conditional density of Y_{it} given (S_{it}, C_t) may not exist. If it does not exist, the probability limit of the kernel estimator may not be measurable with respect to $\sigma(C_t)$, in which case the second step of the estimator, $\widehat{P}_t(s) = h_s(X_t, U_t(s))$, is meaningless.[7]

[7] Formally, the probability limit of kernel estimator can be obtained using the local time of Y_{it}, as in Wang and Phillips (2009). However, the probability limit may not be measurable with respect to $\sigma(C_t)$, as needed here.

We appeal to the disintegration theory for conditional distributions, that can be found in Pollard (2002), to let the common shocks be as general as possible and still having well-defined conditional densities. We therefore impose some restrictions on C_t and work as closely as possible to the standard kernel literature [Souza-Rodrigues (2013)].

For the sake of brevity, we relegate to the appendix (Appendix A.1) the assumptions of the kernel literature adjusted to the present case. Sufficient assumptions to obtain existence are discussed in Souza-Rodrigues (2013) and the references cited there. Briefly, the key sufficient condition is that $\mathcal{C}$ must be a separable metric space equipped with its Borel σ-field. This is satisfied provided the space of functions $\mathcal{J}(\mathcal{S})$ is a separable metric space also equipped with its Borel σ-field, which is a mild restriction. Here we directly impose existence of the conditional density of (Y_{it}^*, S_{it}) given C_t, denoted by $f_t(y^*, s|c)$, where $y^* \in \mathcal{Y}^*$, $s \in \mathcal{S}$ and $c \in \mathcal{C}$.

Remark 2 *To make the conditional density more concrete, take first the example in which S_{it} is binary. In this case, the common shock for group t is the vector $C_t = (X_t, U_{0t}, U_{1t}, Z_t)$ and the conditional density of (Y_{it}^*, S_{it}) given C_t is the usual density*

$$f_{Y_{it}^*,S_{it}|C_t}(y^*, s|c) = f_{Y_{it}^*,S_{it}|X_t,U_{0t},U_{1t},Z_t}(y^*, s|x, u_0, u_1, z).$$

The case where S_{it} is continuous is similar. Take the example where $\mathcal{J}(\mathcal{S})$ is the space $\mathcal{L}^2(\mathcal{S})$. Then $U_t(S) = \sum_{j=1}^{\infty} V_{jt}\psi_j(S)$. The common shock $C_t = (X_t, U_t(\cdot), Z_t)$ can be represented by the infinite dimensional vector $C_t = (X_t, V_{1t}, V_{2t}, ..., Z_t)$. Conditioning on the event $\{C_t = c\}$ is therefore equivalent to conditioning on $\{(X_t, V_{1t}, V_{2t}, ..., Z_t) = (x, v_1, v_2, ..., z)\}$, and so

$$f_{Y_{it}^*,S_{it}|C_t}(y^*, s|c) = f_{Y_{it}^*,S_{it}|X_t,V_{1t},V_{2t},...,Z_t}(y^*, s|x, v_1, v_2, ..., z).$$ [8]

The object of interest in the first step is (for Q_t-almost all $c \in \mathcal{C}$)

$$\begin{aligned} P_t(s,c) &\equiv \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = s, C_t = c) \\ &= \int 1\{y^* \leq D^{-1}(\widetilde{y})\} f_{Y_{it}^*|S_{it},C_t}(y^*|s,c)\, dy^* \\ &= \int_{-\infty}^{D^{-1}(\widetilde{y})} f_{Y_{it}^*|S_{it},X_t,U_t(S_{it})}(y^*|s,x,u(s))\, dy^*. \end{aligned}$$

To use Taylor expansion as usually done in the kernel literature, we need $P_t(s,c)$ to be twice continuously differentiable with respect to s (for Q_t-almost all c). We need to assume the differentiability of $f_{Y_{it}^*|S_{it},X_t,U_t(S_{it})}(y^*|s,x,u(s))$ with respect to $(s, u(\cdot))$ as well as the differentiability

[8] The intuition we may gain in this example, by moving from abstract spaces of functions to space of random vectors, may not apply when $\mathcal{C}$ is not a Hilbert space. The reason is that, although we may approximate any of the separable metric spaces by simpler spaces (such as the space of finite polynomials, for example), the conditioning argument does not hold without running into problems such as the Borel paradox [see, e.g., Rao (1988)].

of $u(\cdot)$ with respect to s. That is why we stated earlier that the main restrictions we need to impose on the space $\left(\mathcal{J}(\mathcal{S}), \|\cdot\|_{\mathcal{J}}\right)$ are: (i) it must be a *separable* metric space equipped with the Borel σ-field (to obtain existence of densities); and (ii) the space $\mathcal{J}(\mathcal{S})$ is a subset of the twice continuously differentiable functions (to obtain the differentiability of $P_t(s,c)$).

The full discussion of the sufficient conditions to obtain consistency with rate is in Appendix A.1. For our purposes here, we only need the following:

Proposition 2 *Let Condition 1 in the main text and Conditions 10-16 in Appendix A.1 hold. Let $\widehat{P}_t(s)$ be the kernel regression estimator defined in (15). Then,*

$$E\left(\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 \mid C_t\right) = e_t(s, C_t) \times N_t^{-\frac{4}{4+d_s}} \tag{16}$$

where d_s is the dimension of S_{it}, and $e_t(s, C_t)$ is a $\sigma(C_t)$-measurable random variable that is Q_t-almost surely finite and is defined by (35) in Appendix A.1.

4.1.3 Uniform Rate

Although Propositions 1 and 2 are necessary for our purposes, they are not sufficient. We need uniform convergence in the first step. More precisely, we need to bound from above the quantity

$$E\left[\max_{1\le t\le T}\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2\right].$$

We obtain this bound in the next Proposition:

Proposition 3 *Let the conditions in Proposition 1 hold if S_{it} is discrete and conditions in Proposition 2 hold if S_{it} is continuous. If, for all $i \geq 1$, $\{Y_{it}, S_{it}, C_t : t \geq 1\}$ are independent across t and if*

$$\sup_{t\ge 1} E\left[e_t(s, C_t)\right] < \infty, \tag{17}$$

then,

$$E\left[\max_{1\le t\le T}\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2\right] \le const. \min\left\{\frac{T}{N_{\min,t}^{2r}}, 1\right\},$$

where $N_{\min,T} \equiv \min\{N_1, ..., N_T\}$; $r = 1/2$ if S_{it} is discrete; and $r = 2/(4+d_s)$ if S_{it} is a d_s-vector of continuously distributed variables.

4.2 Second Step: Across Groups Estimator

In the second step we estimate $h_s(x,u)$. To simplify notation, we denote $P_t(s,C_t)$ by $P_t(s)$. For fixed $s \in \mathcal{S}$ and $u \in (0,1)$, the moment restriction implied by the model is[9]

$$E\left[1\left\{P_t(s) \leq h_s(X_t,u)\right\}|Z_t\right] = 1-u.$$

Define the residual function by

$$\rho_u\left(P_t(s), X_t; h_s\right) \equiv 1\left\{P_t(s) \leq h_s(X_t,u)\right\} - (1-u), \tag{18}$$

then

$$E\left[\rho_u\left(P_t(s), X_t; h_s\right)|Z_t\right] = 0. \tag{19}$$

Denote by h_{0s} the unique (identified) function that satisfies (19). For notational simplicity, sometimes we omit both indices s and u from both $h_s(\cdot,u)$ and $\rho_u(\cdot)$ when it is clear enough from the context. Define the moment function by

$$\begin{aligned} m(Z_t,h_s) &\equiv E\left[\rho_u(P_t(s), X_t; h_s)|Z_t\right] \\ &= E\left[F_{P_s|X,Z}\left(h_s(X_t,u)|X_t,Z_t\right)|Z_t\right] - (1-u) \\ &= \int \left[F_{P_s|X,Z}\left(h_s(x,u)|x,Z_t\right)\right] f_{X|Z}\left(x|Z_t\right)dx - (1-u) \end{aligned} \tag{20}$$

where $F_{P_s|X,Z}$ is the conditional distribution function of $P_t(s)$ given (X_t,Z_t) and $f_{X|Z}$ is the conditional density of X_t given Z_t.

If we observed $P_t(s)$, we could apply the PSMD estimator developed by Chen and Pouzo (2009, 2012) directly. We could let $\overline{m}(Z_t,h_s)$ be an estimator of $m(Z_t,\ h_s)$ (e.g., a series least squares estimator); assume $h_{0s}(x,u)$ belongs to some space of functions $\mathcal{H}$; and let $\{\mathcal{H}_{K(T)} : T = 1,2,...\}$ be an increasing sequence of sieves spaces ($\mathcal{H}_K \subseteq \mathcal{H}_{K+1} \subseteq ... \subseteq \mathcal{H}$) such that $\cup_{T=1}^{\infty}\mathcal{H}_{K(T)}$ is dense in the space $\mathcal{H}$. Then the PSMD estimator $\overline{h}_s \in \mathcal{H}_{K(T)}$ of h_{0s} would minimize the following criterion function:

$$\overline{h}_s = \underset{h_s \in \mathcal{H}_{K(T)}}{\arg\min}\left\{\frac{1}{T}\sum_{t=1}^{T}\overline{m}(Z_t,h_s)'\,\overline{m}(Z_t,h_s) + \lambda_T\widehat{M}_T(h_s)\right\} \tag{21}$$

where the penalization parameter $\lambda_T \geq 0$ is such that $\lambda_T \to 0$ as $T \to \infty$; and $\widehat{M}_T(h)$ is the penalization function. Chen and Pouzo (2012) provide the conditions under which $\overline{h}_s$ is consistent and establish its rate of convergence.

[9] Because $u = \Pr\left[U_t(s) \leq u\right] = \Pr\left[U_t(s) \leq u|Z_t\right] = \Pr\left[h_s\left(X_t,U_t(s)\right) \geq h_s(X_t,u)|Z_t\right] = E\left[1\left\{P_t(s) \geq h_s(X_t,u)\right\}|Z_t\right]$, where the third equality uses the fact that $h_s \downarrow u$.

In the present case however we do not have $P_t(s)$ but rather $\widehat{P}_t(s)$. So we slightly modify the criterion function and use instead

$$\widehat{h}_s = \underset{h_s \in \mathcal{H}_{K(T)}}{\arg\min} \left\{ \frac{1}{T} \sum_{t=1}^{T} \widehat{m}(Z_t, h_s)' \, \widehat{m}(Z_t, h_s) + \lambda_T \widehat{M}_T(h_s) \right\} \tag{22}$$

where $\widehat{m}(Z_t, h_s)$ is an estimator of $m(Z_t, h_s)$ that uses the feasible sample $\left\{ \widehat{P}_t(s), X_t, Z_t \right\}_{t=1}^{T}$. Next, we discuss the consistency of $\widehat{h}_s$ and its convergence rate.

4.2.1 Consistency

We adapt the conditions imposed by Chen and Pouzo (2009, 2012) and assume the following:

Condition 2 *(i)* $\{X_t, U_t(\cdot), Z_t : t \geq 1\}$ *are i.i.d. across* t*, and*

(ii) For any $i, j \geq 1$*, and any* $t \neq t'$*,* (Y_{it}, S_{it}) *is independent of* $\left(Y_{jt'}, S_{jt'}\right)$*.*

Condition 3 *For any* $(y, s, u) \in \mathcal{Y}^* \times \mathcal{S} \times \mathcal{U}$*, assume*

$$\Pr\left(Y_{it}^* \leq y | S_{it} = s, X_t = \cdot, U_t(s) = u\right) \in \mathcal{H}.$$

Define the weight function

$$\omega(x) = \left(1 + \|x\|_E^2\right)^{-\mu/2},$$

for some finite $\mu > 0$*. The parameter space* $\mathcal{H}$ *is either*

$$\mathcal{H} = \{h \in \Lambda^\alpha(\mathcal{X}, \omega) : 0 \leq h \leq 1\},$$

or

$$\mathcal{H} = \left\{h \in W_p^\alpha(\mathcal{X}, \omega) : 0 \leq h \leq 1\right\},$$

where:

(a) $\Lambda^\alpha(\mathcal{X}, \omega)$ *is the weighted Hölder space. Let* $\alpha = \nu + \gamma$*, where* ν *is a nonnegative integer,* $0 < \gamma \leq 1$*, and* $\alpha > 2d_x$*. The weighted Hölder space is*

$$\Lambda^\alpha(\mathcal{X}, \omega) = \{h \in C^\nu(\mathcal{X}) : \|\omega h\|_{\Lambda^\alpha} < \infty\}$$

where $C^\nu(\mathcal{X})$ *is the space of* ν*-times continuosly differentiable functions and*

$$\|h\|_{\Lambda^\alpha} = \max_{0 \leq |\delta| \leq \nu} \left\| D^\delta h \right\|_\infty + \max_{|\delta| = \nu} \sup_{x \neq \bar{x}} \frac{|\; D^\delta h(x) - D^\delta h(\bar{x}) \;|}{\|x - \bar{x}\|_E^\gamma}.$$

(b) $W_p^\alpha(\mathcal{X},\omega)$ *is the weighted Sobolev space. For some scalar integer* $\alpha > 2d_x$*, and for* $1 < p < \infty$*, the weighted Sobolev Space is*

$$W_p^\alpha(\mathcal{X},\omega) = \left\{ h \in L^p(\mathcal{X}, leb) : \|\omega h\|_{W_p^\alpha}^p < \infty \right\},$$

where

$$\|h\|_{W_p^\alpha} = \sum_{0 \leq |\delta| \leq \alpha} \left\| D^\delta h \right\|_{L^p(leb)}.$$

Define the norm $\|.\|_{\mathcal{H}}$ *to be either* $\|h\|_{\mathcal{H}} = \|\omega h\|_{\Lambda^\alpha}$ *if* $\mathcal{H} \subset \Lambda^\alpha(\mathcal{X},\omega)$*, or* $\|h\|_{\mathcal{H}} = \|\omega h\|_{W_p^\alpha}^p$ *if* $\mathcal{H} \subset W_p^\alpha(\mathcal{X},\omega)$*.*

Condition 4 *The penalization parameter is* $\lambda_T \geq 0$*,* $\lambda_T \to 0$ *as* $T \to \infty$*. If* $\lambda_T > 0$*, the penalty function is either*[10]

(a) $\widehat{M}_T(h) = \left\| \left(1 + \|x\|_E^2\right)^{-\zeta/2} h(x) \right\|_{\Lambda^\alpha}$*, with* $\zeta > 0$*, if* $\mathcal{H} \subset \Lambda^\alpha(\mathcal{X},\omega)$*; or*

(b) $\widehat{M}_T(h) = \left\| \left(1 + \|x\|_E^2\right)^{-\zeta/2} h(x) \right\|_{W_p^\alpha}$*, with* $0 < \zeta < \mu$*, if* $\mathcal{H} \subset W_p^\alpha(\mathcal{X},\omega)$*.*

Condition 5 $\{\mathcal{H}_{K(T)} : T = 1, 2, ...\}$ *is a sequence of non-empty closed (under* $\|\cdot\|_{\infty,\omega}$*) subsets of* $\mathcal{H}$ *satisfying* $\mathcal{H}_K \subseteq \mathcal{H}_{K+1} \subseteq \mathcal{H}$ *such that (i)* $\cup_{T=1}^\infty \mathcal{H}_{K(T)}$ *is dense in the space* $\mathcal{H}$ *under* $\|\cdot\|_{\infty,\omega}$*; and (ii)* $\mathcal{H}_{K(T)}$ *is bounded in* $\mathcal{H}$ *under* $\|.\|_{\mathcal{H}}$*.*

Condition 6 $E[\rho_u(P_t(s), X_t; h_{0s})|Z_t] = 0$ *a.s., and for any* $h_s \in \mathcal{H}$ *with* $E[\rho_u(P_t(s), X_t; h_s)|Z_t] = 0$ *a.s., we have* $\|h_s - h_{0s}\|_{\infty,\omega} = 0$*.*

Condition 2(i) requires the common shocks C_t to be i.i.d. across groups t. Because $P_t(s)$ is $\sigma(X_t, U_t(\cdot), Z_t)$-measurable, Condition 2(i) implies that the unfeasible sample $\{P_t(s), X_t, Z_t\}_{t=1}^T$ is i.i.d., as required by Chen and Pouzo (2009, 2012). This condition follows directly from Assumptions A.1, A.2 and Lemma 4 in the Appendix A.3.

Condition 2(ii) imposes independence of (Y_{it}, S_{it}) across groups. It implies that the first step estimators $\left\{\widehat{P}_t(s) : t \geq 1\right\}$ are independent of each other, and, so, the feasible sample $\left\{\widehat{P}_t(s), X_t, Z_t\right\}_{t=1}^T$ is independent across t. However, $\left\{\widehat{P}_t(s), X_t, Z_t\right\}_{t=1}^T$ is not i.i.d. because there is no guarantee that $\left\{\widehat{P}_t(s) : t \geq 1\right\}$ are identically distributed.

[10] Both penalty functions are lower-semicompact. See Edmund and Triebel (1996).

Condition 3 assumes that, for any fixed (y, s, u), $\Pr\left(Y_{it}^* \leq y | S_{it} = s, X_t = \cdot, U_t(s) = u\right)$ belongs to a space that can be well approximated by some sieves spaces. Note that because

$$\Pr\left(Y_{it}^* \leq y \mid S_{it} = s, X_t = x, U_t(s) = u\right) = h_{0s}\left(x^{(1)} + D^{-1}(\widetilde{y}) - y, x^{(2)}, u\right),$$

this condition implies that $h_{0s}(\cdot, u) \in \mathcal{H}$, for any fixed s and u. Because of the large support assumption (Assumption 5), we use the weight function $\omega(.)$ to control the tail behavior of $h_s(.)$.

We obtain consistency of $\widehat{h}_s$ under the weighted *sup-norm*

$$\|h_s\|_{\infty,\omega} = \sup_{x \in \mathcal{X}} |\omega(x) h_s(x, u)|,$$

and we make use of Theorem 3.2 of Chen and Pouzo (2012) because it does not require $\mathcal{H}$ to be compact under the strong norm $\|\cdot\|_{\infty,\omega}$, as in the present case. Theorem 3.2 of Chen and Pouzo (2012) requires lower-semicompact penalty function. I.e., the set $\left\{h \in \mathcal{H} : \widehat{M_T}(h) < M\right\}$, for finite M, must be compact under the norm $\|h_s\|_{\infty,\omega}$. Condition 4 establishes the penalization terms that satisfy these requirements.

Condition 5 allows for both linear and non-linear sieves spaces, $\mathcal{H}_{K(T)}$. It is satisfied, in particular, if we use

$$\begin{aligned} \mathcal{H}_{K_T} &= \left\{h \in \mathcal{H} : h(\cdot) = \sum_{k=1}^{K_T} a_k \psi_k(\cdot), \ 0 \leq h \leq 1, \|h\|_{\mathcal{H}} \leq B_T\right\}, \\ B_T &\rightarrow \infty \text{ slowly as } T \rightarrow \infty, \end{aligned}$$

where $\{\psi_k\}_{k=1}^{k_h}$ is a sequence of known tensor product basis functions such as splines and Fourier series if $\mathcal{H} \subset \Lambda^\alpha(\mathcal{X}, \omega)$, and wavelets if $\mathcal{H} \subset W_p^\alpha(\mathcal{X}, \omega)$. In these cases $\mathcal{H}_{K(T)}$ is compact in $\mathcal{H}$ under $\|\cdot\|_{\infty,\omega}$ and bounded under $\|.\|_{\mathcal{H}}$. Shape-preserving sieves that nonlinearly transform the linear approximations also can be used. Condition 5(ii) is not imposed by Chen and Pouzo (2012), but is important here to control how fast K_T can increase with T.

Condition 6 imposes identification of h_{0s}. It follows from Assumptions 1-7 and Theorems 1 and 2, as discussed in Section 3.

Given the definition of the sieves estimator (equation (22)), the assumptions on the feasible (and unfeasible) data set (Condition 2), the parameter space (Condition 3), the penalization term (Condition 4), the approximating sieves spaces (Condition 5) and the identification result (Condition 6), all we need next is the uniform consistency of the criterion function in order to prove consistency of the estimator. This requires some restrictions on the estimator $\widehat{m}(Z_t, h)$ and on the moment function $m(Z_t, h)$.

We take $\widehat{m}(Z_t, h)$ to be a series least square estimator of $m(Z_t, h)$. So, let $\{p_1(Z), p_2(Z), ...\}$ be a sequence of known basis functions that approximate any square integrable real-valued function of $\mathcal{Z}$ well. Denote $p^{J_T}(Z) = (p_1(Z), ..., p_{J_T}(Z))'$ a $(J_T \times 1)$-vector and $P = (p^{J_T}(z_1)', ..., p^{J_T}(z_T)')'$ a $(T \times J_T)$-matrix. The series least squares estimator is given by:

$$\widehat{m}(z, h_s) = p^{J_T}(z)'(P'P)^{-} \sum_{t=1}^{T} p^{J_T}(Z_t)\rho_u\left(\widehat{P}_t(s), X_t; h_s\right)$$

where $(P'P)^{-}$ is the pseudo-inverse matrix of $P'P$. We impose the following:

Condition 7 *(i) $\mathcal{Z}$ is a compact connected subset of $\mathbb{R}^{d_z}$ with Lipschitz continuous boundary and f_Z is bounded and bounded away from zero over $\mathcal{Z}$;*

(ii) $\max_{1\leq j\leq J_T} E\left[|p_j(Z_t)|^4\right] \leq$ const.; and the smallest and largest eigenvalues of $E\left[p^{J_T}(Z)p^{J_T}(Z)'\right]$ are bounded and bounded away from zero for all J_T;

(iii) Denote $\xi_T \equiv \sup_{z\in\mathcal{Z}} \|p^{J_T}(z)\|_E$, and let $\xi_T^2 J_T = o(T)$;

(iv) There exists of a function $p^{J_T}(Z)'\pi$ such that, uniformly over $h \in \mathcal{H}_{K(T)}$,

$$\left\|m(\cdot, h_s) - p^{J_T}(\cdot)'\pi\right\|_{L_Q}^2 = O_p\left(b_{m,J_T}^2\right),$$

where L_Q is either the L_2 or the L_∞ norm.

Condition 8 *(i) For each $s \in \mathcal{S}$, $F_{P_s|X,Z}$ has a density, $f_{P_s|X,Z}(p|x,z)$, that is continuous in (p, x, z) and bounded, $\sup_{p\in[0,1]} f_{P_s|X,Z}(p|x,z) \leq K < \infty$, for some finite K and for almost all (x, z); and (ii) for some $\kappa > \alpha - \frac{d_x}{2} > 0$, $E\left[\left(1 + \|X_t\|_E^2\right)^{2(\mu+\kappa)}\right] < \infty$.*

Condition 7 is imposed in Chen and Pouzo (2012) and is standard in the nonparametric series regression literature, except that we require existence of higher order moments for $E\left[|p_j(Z_t)|^4\right]$. If $p^{J_T}(Z)$ is a spline, cosine/sine or wavelet sieve, then $\xi_T^2 \asymp J_T$; see e.g. Newey (1997) or Huang (1998). In this case Condition 7(iii) requires $J_T^2/T = o(1)$. Condition 7(iv) is satisfied with $b_{m,J_T} = J_T^{-\alpha_m/d_z}$ if $m(\cdot, h_s)$ belongs to a Hölder space: $m(\cdot, h_s) \in \Lambda^{\alpha_m}(\mathcal{Z})$, with $\alpha_m > d_z/2$ for all $h \in \mathcal{H}_{K_T}$.

Finally, Condition 8 implies that $E\left\{[m(Z_t, h_s)]^2\right\}$ is continuous on $\left(\mathcal{H}, \|\cdot\|_{\infty,\omega}\right)$ and Condition 8(ii) imposes existence of higher order moments of X_t.

Define $N_{\min,T} \equiv \min\{N_1, ..., N_T\}$. Proposition 4 follows:

Proposition 4 *Let Conditions 2-8 hold. Also, let the conditions in Proposition 3 hold for the first step estimator. Let (i) $J_T, K_T \to \infty$, $K_T \asymp J_T$, $K_T \leq J_T$, (ii) $\xi_T^2 \asymp J_T$, (iii)*

$$\left[\frac{K_T}{T} + \frac{K_T^2}{T}\min\left\{\frac{T}{N_{\min,T}^{2r}}, 1\right\}\right] \to 0 \tag{23}$$

as $(T, N_{\min,T}) \to \infty$, *and (iv)*

$$\max\left\{\left(\frac{K_T}{T}+\frac{K_T^2}{T}\min\left\{\frac{T}{N_{\min,T}^{2r}},1\right\}\right),\left(\frac{J_T}{T}+b_{m,J_T}^2\right)\right\}=O\left(\lambda_T\right).$$

Then, for a fixed $s \in \mathcal{S}$ *and a fixed* $u \in (0,1)$*,*

$$\left\|\widehat{h}_s\left(u\right)-h_{0s}(u)\right\|_{\infty,\omega}=o_p\left(1\right)$$

which implies

$$\left\|\widehat{h}_s(u)-h_{0s}(u)\right\|_{L_2(f_X)}=o_p\left(1\right)$$

Remark 3 *The proof here is similar to Chen and Pouzo (2012), except that, by considering the first step estimator, we have to impose both the extra condition (23) and the results of Proposition 3. The extra conditions are needed to guarantee that the feasible criterion function is uniformly close (with probability approaching 1) to the populational criterion function. The main difficulty in the proof comes from the fact that the preliminary estimator breaks the* L^2*-Lipschitz condition of the criterion function used by Chen and Pouzo (2012). Their proof of consistency relies on an empirical process argument in which the Lipschitz condition is used to bound the entropy number of the criterion function and, so, to uniformly approximate it to its populational version. In the present case, however, we cannot rely on such argument, so we have to check explicitly how the entropy number depends on the behavior of the first step estimator and impose extra conditions to control its behavior.*

Remark 4 *From equation (12), it is clear that we could rewrite the model as (see Horowitz and Lee, 2009)*

$$P_t\left(s\right)=h_s\left(X_t\right)+U_t\left(s\right),\ \textit{with}\ \Pr\left[U_t\left(s\right)\leq u|Z_t\right]=u.$$

In the present case we can view the model as

$$P_t\left(s\right)=h_s\left(X_t\right)+U_t\left(s\right)+\varepsilon_t\left(s\right),\ \textit{with}\ \Pr\left[U_t\left(s\right)\leq u|Z_t\right]=u,$$

where $\varepsilon_t\left(s\right)=P_t\left(s\right)-\widehat{P}_t\left(s\right)$ *is the estimation error from the first step. The composite error here* $U_t\left(s\right)+\varepsilon_t\left(s\right)$ *has larger variance than the error term in the model where* $P_t\left(s\right)$ *is known. Intuitively, to control the variance of* $\widehat{h}_s$ *we need to control the variance of the errors. Because the criterion function is non-linear in* $\varepsilon_t\left(s\right)$ *we cannot average these errors out. So we need to control the variance of* $\max_{1\leq t\leq T}\left\{\varepsilon_t\left(s\right)\right\}$*. On the one hand, for any group* t*,* $\varepsilon_t\left(s\right)\xrightarrow{p}0$ *as* $N_t\to\infty$*, and* $N_t^{2r}\varepsilon_t\left(s\right)$ *converges in distribution to a mixing normal, where the mixing depends on the* *common shocks C_t (Andrews, 2005; Souza-Rodrigues, 2013). But for any fixed N_t's, the variance of* $\max_{1\leq t\leq T}\left\{N_t^{2r}\varepsilon_t\left(s\right)\right\}$ *increases as $T\rightarrow\infty$. The contribution of $\varepsilon_t\left(s\right)$ to the variance of $\widehat{h}_s$ depends therefore on how fast $T\rightarrow\infty$ compared to $N_{\min,T}\rightarrow\infty$. From the proof of Proposition 4 it is clear that we need K_T to satisfy the restriction (23) to control the variance of the errors. Note that consistency is achieved even if $\frac{T}{N_{\min,T}^{2r}}\rightarrow\infty$, provided $K_T^2/T\rightarrow 0$ as $(T,N_{\min,T})\rightarrow\infty$. I.e., even if there are many more groups T than individuals in each group N_t.*

4.2.2 Rate of Convergence

Next we obtain the rate of convergence for $\widehat{h}_s$. Following the arguments in Chen and Pouzo (2009, 2012), we derive the rate under a weak norm first, $\left\|\cdot\right\|$, and then derive the rate for the stronger norm. To obtain explicit rates in terms of the number of observations, we restrict the parameter space to be in a Hilbert space, $\mathcal{H}\subset W_2^{\alpha}\left(\mathcal{X},\omega\right)$. We take the strong norm to be the $L^2\left(f_X\right)$ norm, $\left\|\cdot\right\|_{L_2\left(f_X\right)}$, and we restrict the sieves space accordingly: assume it is a tensor product basis of the $L^2\left(f_X\right)$ space. Note that wavelets form an appropriate basis for $L^2\left(f_X\right)$.

With an appropriate basis for the $L^2\left(f_X\right)$ space it is possible to link the weak and the strong norms and obtain explicit rates of convergence in terms of the number of groups and the number of individuals within each group. The term that links both norms is the function $\varphi\left(.\right)$ defined in Condition 19(ii) in Appendix A.2. The functional form of φ depends upon whether we have a mildly or a severely ill-posed problem.

Define

$$\begin{aligned} r_{NT} &= \max\left\{\frac{K_T}{T},\frac{K_T^2}{T}\min\left\{\frac{T}{N_{\min,T}^{2r}},1\right\}\right\}, \textit{and} \\ \delta_{NT}^2 &= \max\left\{r_{NT},\frac{J_T}{T},b_{m,J_T}^2\right\}. \end{aligned}$$

Inspecting the proofs of the Proposition 4 above, and both Chen and Pouzo (2012)'s Corollary 5.1 and Proposition 6.2, it is clear that the only difference between our case and theirs is that our term δ_{NT} includes the extra r_{NT} because of the first step estimator. But other than that, Chen and Pouzo's (2012) results are directly applicable. For this reason we briefly state the additional conditions required to obtain the result in Appendix A.2 and present the rate of convergence in the Proposition 5 below without proof.

The cases to consider depend on whether we have: (a) a mildly or severely ill-posed problem and (b) which term dominates r_{NT}.

Proposition 5 *Let the conditions stated in Proposition 4 (with Conditions 3 and 5 replaced by Condition 18 in Appendix A.2 where appropriate), and Conditions 18-19 in Appendix A.2 hold. Let $\varphi(\cdot)$ be an increasing function defined in Condition 19(ii) (in Appendix A.2). If* $\max\left\{r_{NT}, \frac{J_T}{T}, b^2_{m,J_T}, \lambda_T\right\} = r_{NT}$*, then*

$$\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left(K_T^{-\alpha/d_x} + \sqrt{\frac{r_{NT}}{\varphi\left(K_T^{-2/d_x}\right)}}\right).$$

1. ***Mildly ill-posed problem****: let $\varphi(\tau) = \tau^{\varsigma}$, for some $\varsigma \geq 0$.*
   (a) ***"Large-N" Case:*** *If $\left(\frac{T}{N^{2r}_{\min,T}}\right) \to 0$ as $(T, N_{\min,T}) \to \infty$, then*

   $$\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left(T^{-\frac{\alpha}{2(a+\varsigma)+d_x}}\right),$$

   provided $K_T \asymp T^{\frac{d_x}{2(a+\varsigma)+d_x}}$ and

   $$\left(\frac{T}{N^{2r}_{\min,T}}\right)\left[T^{\frac{d_x}{2(a+\varsigma)+d_x}}\right] \to c < \infty$$

   as $(T, N_{\min,T}) \to \infty$, for some $c > 0$.

   (b) ***"Large-T" Case:*** *If $\left(\frac{T}{N^{2r}_{\min,T}}\right) \to \overline{c} > 0$ as $(T, N_{\min,T}) \to \infty$, then*

   $$\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left(T^{-\frac{\alpha}{2(a+\varsigma)+2d_x}}\right).$$

   provided $K_T \asymp T^{\frac{d_x}{2(a+\varsigma)+2d_x}}$.

2. ***Severely ill-posed problem****: let $\varphi(\tau) = \exp\left(-\tau^{-\varsigma/2}\right)$, for some $\varsigma > 0$.*
   (a) ***"Large-N" Case:*** *If $\left(\frac{T}{N^{2r}_{\min,T}}\right) \to 0$ as $(T, N_{\min,T}) \to \infty$, then*

   $$\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left([\ln T]^{-\frac{\alpha}{\varsigma}}\right),$$

   provided $K_T \asymp [\ln T]^{\frac{d_x}{\varsigma}}$ and

   $$\left(\frac{T}{N^{2r}_{\min,T}}\right)[\ln T]^{\frac{d_x}{\varsigma}} \to c < \infty$$

   as $(T, N_{\min,T}) \to \infty$, for some $c > 0$.

   (b) ***"Large-T" Case:*** *If $\left(\frac{T}{N^{2r}_{\min,T}}\right) \to \overline{c} > 0$ as $(T, N_{\min,T}) \to \infty$ then*

   $$\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left([\ln T]^{-\frac{\alpha}{\varsigma}}\right),$$

   provided $K_T \asymp [\ln T]^{\frac{d_x}{\varsigma}}$.

Remark 5 ***(Mildly ill-posed problem, "Large-N" case).*** *If $P_t(s)$ were known, the rate of convergence in the mildly ill-posed problem would be $\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left(T^{-\frac{\alpha}{2(a+\varsigma)+d_x}}\right)$, as showed by Chen and Pouzo (2012). In the present case however the rate depends on how fast $N_{\min,T} \to \infty$ compared to $T \to \infty$. In the "Large-N"case, we have $\left(\frac{T}{N_{\min,T}^{2r}}\right) \to 0$, as $(T, N_{\min,T}) \to \infty$. Because the number of individuals in each group is much larger than the number of groups, the first step does not affect the convergence rate. The resulting rate is the same rate we would obtain if we observed $P_t(s)$.*[11]

Remark 6 ***(Mildly ill-posed problem, "Large-T" case).*** *When the number of individuals in each group is smaller (or not much larger) than the number of groups, the rate of convergence is slower compared to the "Large-N" case: $\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left(T^{-\frac{\alpha}{2(a+\varsigma)+2d_x}}\right)$, instead of $\left\|\widehat{h}_s(u) - h_{0s}(u)\right\|_{L_2(f_X)} = O_p\left(T^{-\frac{\alpha}{2(a+\varsigma)+d_x}}\right)$. We define the "Large-T" case when $\left(\frac{T}{N_{\min,T}^{2r}}\right) \to \overline{c} > 0$ as $(T, N_{\min,T}) \to \infty$. If $r = 1/2$ (in case S_{it} is discrete), we can have $T \simeq \overline{c}N_{\min,T}$ either for $0 < \overline{c} < 1$ (T smaller than $N_{\min,T}$) or for $\overline{c} > 1$ (T is larger than $N_{\min,T}$).*

The variance of the errors $\max_{1\leq t\leq T}\left\{P_t(s) - \widehat{P}_t(s)\right\}$ in the "Large-T" case does not die out as fast as in the "Large-N" case. To control that variance we need therefore to choose K_T that increases slower than in the previous case. So, by taking $K_T \asymp T^{\frac{d_x}{2(a+\varsigma)+2d_x}} \lesssim T^{\frac{d_x}{2(a+\varsigma)+d_x}}$, we control the variance of $\widehat{h}_s$ and obtain the rate $O_p\left(T^{-\frac{\alpha}{2(a+\varsigma)+2d_x}}\right)$. The rate is almost the same as the optimal rate, except for the extra factor 2 multiplying d_x, which slows down the rate of convergence compared to the optimal rate.[12]

Remark 7 ***(Severely ill-posed problem).*** *For the severely ill-posed case, the rate of convergence is sufficiently slow even if we observed $P_t(s)$. As a result it is not affected by the first step estimator regardless how fast $N_{\min,T} \to \infty$ compared to $T \to \infty$.*[13]

Remark 8 *Despite the fact $h_{0s}(\cdot, u)$ is strictly decreasing in u by Assumption 1, $\widehat{h}_s(\cdot, u)$ is not*

[11] Note that if $\left(\frac{T}{N_{\min,T}^{2r}}\right) \to 0$, then $r_{NT} = \frac{K_T}{T}\max\left\{1, K_T\frac{T}{N_{\min,T}^{2r}}\right\}$, for large $(K_T, T, N_{\min,T})$. So, if we take K_T such that $K_T\frac{T}{N_{\min,T}^{2r}} \to c < \infty$, then $r_{NT} = \frac{K_T}{T}\overline{c}$, with $\overline{c} \geq 1$. The choice of K_T that balances bias and variance satisfies $K_T^{-2\alpha/d_x}\varphi\left(K_T^{-2/d_x}\right) \asymp r_{NT}$. In the mildly ill-posed problem, $K_T \asymp T^{\frac{d_x}{2(a+\varsigma)+d_x}}$ balances bias and variance, provided $\left(\frac{T}{N_{\min,T}^{2r}}\right)\left[T^{\frac{d_x}{2(a+\varsigma)+d_x}}\right] \to c < \infty$.

[12] If $\left(\frac{T}{N_{\min,T}^{2r}}\right) \to \overline{c} > 0$, then $r_{NT} = \frac{K_T}{T}\max\left\{1, K_T\overline{\overline{c}}\right\}$, for $\overline{\overline{c}} \leq 1$, for large $(T, N_{\min,T})$. Eventually $K_T\overline{\overline{c}} > 1$, implying $r_{NT} = \frac{K_T^2}{T}\overline{\overline{c}}$. The choice of K_T that balances bias and variance for the mildly ill-posed problem is then $K_T \asymp T^{\frac{d_x}{2(a+\varsigma)+2d_x}}$.

[13] The arguments for obtaining the K_T that balances bias and variance for the severely ill-posed problem are the same as the arguments presented before for the midly ill-posed problems.

guaranteed to be strictly decreasing. One may need to make adjustments in the spirit of Chernozhukov, Fernandez-Val and Galichon (2010) to guarantee the monotonicity of $\widehat{h}_s(\cdot,u)$ *in* u*.*

We conclude that the preliminary estimator $\widehat{P}_t(s)$ may or may not affect the rate of convergence of the second step estimator. It depends upon (a) whether r_{NT} dominates the term δ_{NT}; (b) whether we have a mildly ill-posed or a severely ill-posed problem; and (c) the rate at which $(T, N_{\min,T}) \to \infty$.

5 Monte Carlo Simulation

[TO BE FINISHED]

We report a small Monte Carlo (MC) study for the model:

$$\begin{aligned} Y_{it} &= 1\left[Y_{it}^{*} \geq 0\right] \\ Y_{it}^{*} &= X_t^{(1)} + g\left(S_{it}, X_t^{(2)}, U_t\left(S_{it}\right), \varepsilon_{it}\right). \end{aligned} \tag{24}$$

where

$$g\left(.\right) = \alpha + \beta S_{it} + \gamma_1 X_t^{(2)} + \gamma_2 X_t^{(2)} \varepsilon_{it} + \Phi\left(U_t\left(S_{it}\right)\right) + \eta_{it},$$

where Φ is the cumulative distribution function of the standard normal. Let $(S_{it}, \varepsilon_{it}, \eta_{it}) \sim N(0, I)$, with I the identity matrix; and let the special regressor be $X_t^{(1)} \sim N(0,1)$. The endogenous group-level variable is given by

$$X_t^{(2)} = \pi Z_t + V_{0t}$$

where π measures the strength of the instrument Z_t, and V_{0t} is correlated with the group-level unobservable $U_t(.)$. Because S_{it} is continuously distributed, we let $U_t(.) \in \mathcal{J}(\mathcal{S}) \subseteq \mathcal{L}^2(\mathcal{S})$, and take a basis $\left\{\psi_j\right\}_{j=1}^{\infty}$ so that $U_t(S) = \sum_{j=1}^{\infty} V_{jt}\psi_j(S)$. In practice we take $\left\{\psi_j\right\}_{j=1}^{J}$ to be a $J = 4$ finite-order polynomial. The combined vector of V_{0t} and the coefficients of $U_t(.)$ has a joint normal distribution: $(V_0, V_1, ..., V_J) \sim N(0, \Omega)$, with a non-diagonal variance-covariance matrix Ω.

Because the model is the binary choice model, we take $\widetilde{y} = 0$, and so $D^{-1}(\widetilde{y}) = 0$. The object of interest is[14]

$$\Pr\left(Y_{it}^{*} \leq y \mid S_{it} = s, X_t^{(1)} = x^{(1)}, X_t^{(2)} = x^{(2)}, U_t(s) = u\right) = h_s\left(x^{(1)} - y, x^{(2)}, u\right).$$

[14] As a matter of fact,

$$\begin{aligned} & \Pr\left(Y_{it}^{*} \leq y \mid S_{it} = s, X_t = x, U_t(s) = u\right) \\ = & \int 1\left\{x^{(1)} + \delta\Phi\left(\frac{1}{\sigma}\left[\alpha + \beta s + \gamma_1 x^{(2)} + \gamma_2 x^{(2)}\varepsilon_{it} + \Phi(u) + \eta_{it}\right]\right) \leq y\right\} dF\left(\varepsilon_{it}, \eta_{it}\right). \end{aligned}$$

This conditional probability is a smooth function of $\left(s, x^{(1)}, x^{(2)}, u\right)$.

For the first step, we estimate $P_t(s)$ for each group t using the standard kernel regression and the "leave-one-out" cross-validation method to select the bandwidth.

For the second step, we use the shape-preserving sieves $\mathcal{H}_{K_T}$... The moment funtion is approximated by cubic splines; the penalization $M_T(h)$ is...; and the tuning parameters are (K_T, J_T, λ_T).

We run 500 Monte Carlo repetitions for each (N,T) ranging from $(250,500)$ to $(1500,1000)$. Table XX shows the integrated square bias (I-BIAS^2), the integrated variance (I-VAR), and the integrated mean square error (I-MSE). The results are...

6 Empirical Application

[TO BE FINISHED]

As discussed in Example 2, the impact of hospital volumes of surgical procedures on individual mortality rates is a question of public interest. Numerous studies have documented an inverse relationship between hospital volumes and mortality rates after surgery. But the evidence is weak for most operations [Birkmeyer et al. (2002), and Finks, Osborne and Birkmeyer (2011)]. Given the evidence for some selected operations, there has been a considerable interest in the United States since the early 2000s in concentrating the procedures in high-volume hospitals. Indeed, the Leapfrog Group, a large coalition of private and public purchasers of health insurance, has been encouraging volume-based referral in the last decade. The objective is to reduce the number of preventable deaths in surgical procedures performed by hospitals with inadequate experience (i.e., with low volume of operations). In the presence of economies of scale or learning, concentrating operations may be welfare increasing.

We collected data from several sources. The Medicare Provider Analysis and Review (MEDPAR) provided the Limited Data Set (LDS) Denominator files and the LDS Inpatient Standard Analytic files. The LDS Denominator files contain the characteristics of the patients, including the date of death after the surgery (if ocurred before hospital discharged); the LDS Inpatient Standard Analytic files contain information on the date of the surgeries and the type of surgery (from the International Classification of Diseases Code – the ICD-9 code). These data were combined with the State Inpatient Database (SID), which contains hospitals characteristics, including the volumes of operations per hospital per surgery and per year. Because the SID does not provide data for all states and years, we selected the years from XXXX to XXXX for the states listed in the Appendix XX.[15] To make the results comparable to Birkmeyer et al. (2002), we selected the cardiovascular

[15] We linked the data sets using the National Provider Identifier (NPI) and the Medicare Provider Number (that has been renamed the CMS Certification Number – CCN) in the LDS files and the American Health Association

and cancer resection surgeries. We limited the data to patients who were over 65 years of age and under 99 years of age.

The model we take to the data is

$$\begin{aligned} Y_{it} &= D\left(Y_{it}^{*}\right) \\ Y_{it}^{*} &= X_t^{(1)} + g\left(S_{it}, X_t^{(2)}, U_t\left(S_{it}\right), \varepsilon_{it}\right). \end{aligned} \tag{25}$$

where i is the index for patients, and t, for hospitals. The latent Y_{it}^{*} represents a continuous measure of actual health status; and the observable outcome Y_{it} is the binary indicator that equals 1 if the patient died before hospital discharged or within 30 days after the surgical procedure, and equals 0 otherwise. The vector S_{it} includes patients' gender (male/female) and age group (65 to 74, 75 to 84, and 85 to 99 years). The vector of hospital characteristics, X_t, includes the size of the hospital, say, the number of beds ($X_t^{(1)}$) and the volume of operations ($X_t^{(2)}$). The unobserved hospital quality, $U_t\left(.\right)$ is indexed by S_{it} because an unobserved hospital characteristic that is helpful for patients with one S_{it} may be harmful to other patients. Because hospitals with high volumes of surgeries are likely those with high unmeasured quality, we make use of instrumental variables. The IV in this application is the number of hospital per-capita in the region. The number of hospital per-capita should affect the level of local competition, and so affect the equilibrium volumes of surgeries, but should be excluded from the individual mortality equation.

We estimate the conditional probability of survival $\Pr\left(Y_{it} = 0 | S_{it}, X_t, U_t\left(S_{it}\right)\right)$. From Section 3, we know that

$$\begin{aligned} & \Pr\left(Y_{it} = 0 \mid S_{it} = s, X_t = x, U_t\left(s\right) = u\right) \\ = \; & \Pr\left(Y_{it}^{*} \leq D^{-1}\left(0\right) \mid S_{it} = s, X_t = x, U_t\left(s\right) = u\right) \\ = \; & h_s\left(x^{(1)}, x^{(2)}, u\right). \end{aligned}$$

because we must take $\widetilde{y} = 0$ when the endogenous variable is binary. Next, discuss the specific choices for sieves spaces and penalization.

Results. Compare (a) logit; (b) logit with IV; (c) nonparametric model. The results are...

7 Conclusion

This paper develops a nonparametric estimator for the generalized regression model proposed by Berry and Haile (2009) in which each individual is associated with a group and each group is

(AHA) hospital identifier in the SID. The AHA database provides all hospital identifiers (NPI, CCN and AHA ID) to allow the merger of the LDS files with the SID.

subject to observable and unobservable shocks. The motivation for this model is to estimate the effects of group-level observables on individual outcomes when group-level observables correlate with group-level unobservables. The group-level unobservables are allowed to be indexed by individual characteristics, which allows for more general group effects than existing approaches. In the binary choice demand example, indexing this unobservable might capture how men and women rank differently the unobserved quality of the good. Their ranks could be completely different and independent of each other, a possibility that is not allowed if we restrict the group-level unobservables to be a scalar random variable. The same reasoning applies in case the individual characteristics are continuous, such as income, for example.

We propose a two-step estimator in which the first step runs a nonparametric regression of individual outcomes on individual observables within each group. It is a nonparametric regression in the presence of common shocks. The second step fixes the individual characteristics and runs a nonparametric quantile instrumental variable regression across groups of the predicted outcome obtained in the first step on group-level variables. It separates the effects of group-level observables from unobservables. The second step modifies the penalized sieve minimum distance estimator (PSMD) developed by Chen and Pouzo (2009, 2012) to take into account the preliminary estimator from the first step. This paper establishes sufficient conditions to obtain consistency and convergence rate of the estimator. The rate of convergence depends on the rates at which both the number of groups and the number of observations within each group go to infinity.

We plan to investigate the properties of the asymptotic distribution of the two-step estimator in the future.

References

[1] Abrevaya, J. (2000). “Rank Estimation of a Generalized Fixed Effects Regression Model,” *Journal of Econometrics*, 95, 1–23.

[2] Ai, C. and X. Chen (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica* 71 1795-1844.

[3] Altonji, J., and R. L. Matzkin (2005): “Cross-Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors,” *Econometrica*, 73, 1053–1102.

[4] Andrews, D. W. K. (1994). "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Volume 4, ed. by R.F. Engle and D. McFadden. New York: North-Holland, 1994, 2247-2294.

[5] Andrews, D. W. K. (2005). "Cross-section Regression with Common Shocks," *Econometrica*, 73, 1551–1585.

[6] Berry, S. T. and P. A. Haile (2009). "Identification of a Nonparametric Generalized Regression Model with Group Effects," Discussion paper, Yale University.

[7] Berry, S. T., and P. A. Haile (2013). "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*,

[8] Berry, S. T., and P. A. Haile (2010): "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," Discussion paper, Yale University.

[9] Birkmeyer et al. (2002)

[10] Chen, X. (2007). "Large Sample Sieve Estimation of Semi-nonparametric Models," chapter 76 in *Handbook of Econometrics*, Vol. 6B, 2007, eds. James J. Heckman and Edward E. Leamer, North-Holland.

[11] Chen, X. and D. Pouzo (2008). "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals" Cowles Foundation Discussion Paper No. 1650

[12] Chen, X. and D. Pouzo (2009) "Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals", *Journal of Econometrics*, vol. 152, pp. 46-60.

[13] Chen, X. and D. Pouzo (2012). "Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals," *Econometrica*, 80, No. 1, 277-321.

[14] Chen, X., Chernozhukov, V. Lee, S. and W. K. Newey (2011). "Identification in Semiparametric and Nonparametric Conditional Moment Models," Discussion Paper 1795, Coweles Foundation.

[15] Chernozhukov, V. and C. Hansen (2005). "An IV Model of Quantile Treatment Effects," *Econometrica* 73, 245–261.

[16] Chernozhukov,V. and C. Hansen (2008). "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, Volume 142, Issue 1, 379-398.

[17] Chernozhukov, V., I. Fernández-Val and A. Galichon (2010). "Quantile and Probability Curves Without Crossing," *Econometrica*, 78, Issue 3, 1093-1125.

[18] Chernozhukov, V., G. Imbens, and W. Newey (2007). "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4-14.

[19] Chiappori, P. A., I. Komunjer and D. Kristensen (2011). "Correct Specification and Identification of Nonparametric Transformation Models," Discussion paper, University of California San Diego.

[20] Durlauf, S. N., S. Navarro and D. A. Rivers (2010). "Understanding Aggregate Crime Regressions," *Journal of Econometrics*, 158, Issue 2, 306-317.

[21] Evdokimov, K. (2009). "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity", Yale University, Working paper.

[22] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, Vol. II (2nd ed.). New York: Wiley.

[23] Finks, Osborne and Birkmeyer (2011)

[24] Fox, J., and A. Gandhi (2011). "Identifying Demand with Multidimensional Unobservables: A Random Functions Approach," Working Paper.

[25] Han, A. K. (1987). "Nonparametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.

[26] Hoderlein,S. and H. White: "Nonparametric Identification in Nonseparable Panel Data Models with Generalized Fixed Effects," Working Paper

[27] Honoré, B. E., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," *emet*, 70(5), 2053–2063.

[28] Horowitz, J. L., and S. Lee (2007). "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75, 1191–1208.

[29] Huang, J. (1998). "Projection Estimation in Multiple Regression with Application to Functional ANOVA Models," *Annals of Statistics*, 26, 242-272.

[30] Ichimura, H., and T. S. Thompson (1998): “Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution,” *Journal of Econometrics*, 86(2), 269–95.

[31] Manski, C. F. (1985). “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,” *Journal of Econometrics*, 27, 313–333.

[32] Meyer, P. A. (1966). *Probability and Potentials*. Blaisdell Publishing Co, New York.

[33] Newey, W. K. (1997). "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147-168.

[34] Nickl, R. and B. M. Pöscther (2007). "Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type," *Journal of Theoretical Probability*, 20, 177-199.

[35] Phillips, P. C. B. and H. R. Moon (1999). "Linear Regression Limit Theory for Nonstationary Panel Data" *Econometrica*, Vol. 67, No. 5, 1057–1111.

[36] Pollard, D. (2002). *A User’s Guide to Measure Theoretic Probability.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[37] Pollard, D. (2002). "Maximal Inequalities Via Bracketing with Adaptive Truncation," *Annales de l’Institut Henri Poincare (B) Probability and Statistics*, Volume 38, Issue 6, 1039-1052.

[38] Souza-Rodrigues, E. A. (2014a). "Demand for Deforestation in the Amazon," Yale University.

[39] Souza-Rodrigues, E. A. (2014b). "Nonparametric Regression with Common Shocks," Yale University.

[40] Su, L., S. Hoderlein, and H. White (2010). "Testing Monotonicity in Unobservables with Panel Data," Cowles Conference, June 2010.

[41] Torgovitsky, A. (2011). "Identification of Nonseparable Models with General Instruments," Working Paper, Yale University.

[42] Van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*, New York: Springer-Verlag.

[43] Wang, Q. and P. C. B. Phillips (2009). "Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegration Regression," *Econometric Theory*, 25, 710–738.

A Appendix

The Appendix is divided as follows. First we present the proofs for the results of both the first and the second steps of the estimator (Appendix A.1 and A.2, respectively). Then we describe the probabilistic framework, based on Andrews (2005), that justifies the approach taken in this paper (Appendix A.3).

A.1 First Step - Within Groups

In this subsection we present the results for the first step estimator. We consider two cases: S_{it} is discrete and S_{it} is continuous.

A.1.1 Discrete Case

Following the main text, suppose S_{it} can take finitely many values $\{0, 1, .., L\}$. Define for $l = 1, .., L$,

$$S_{it}^{l} = \begin{cases} 0 \text{ if } S_{it} \neq l \\ 1 \text{ if } S_{it} = l \end{cases}$$

and define $\Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it}^{1}, ..., S_{it}^{L}, C_t\right)$ by

$$\Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it}^{1}, ..., S_{it}^{L}, C_t\right) = \beta_0\left(C_t\right) + \beta_1\left(C_t\right) \times S_{it}^{1} + ... + \beta_L\left(C_t\right) \times S_{it}^{L}.$$

As a result, we have that

$$\begin{aligned} \Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it} = 0, C_t\right) &= \beta_0\left(C_t\right) \\ \Pr\left(Y_{it} \leq \widetilde{y} \mid S_{it} = l, C_t\right) &= \beta_0\left(C_t\right) + \beta_l\left(C_t\right), \end{aligned}$$

for $l = 1, .., L$. For each group t, we run a linear regression of $1\{Y_{it} \leq \widetilde{y}\}$ on $S_{it}^{1}, ..., S_{it}^{L}$ and take the estimated prediction of $1\{Y_{it} \leq \widetilde{y}\}$ given $S_{it} = s$ as our estimator $\widehat{P}_t(s)$, i.e., $\widehat{P}_t(s) = \widehat{\beta}_0 + \widehat{\beta}_s$.

In order to apply the asymptotic results in Andrews (2005), we need to guarantee his Assumptions 1-3. Assumption 1 is obtained directly from Condition 1 in the main text. The other assumptions (about existence of moments) are satisfied – except for Assumption 2(d) – because we are regressing a binary variable $1\{Y_{it} \leq \widetilde{y}\}$ on other binary variables $S_{it}^{1}, ..., S_{it}^{L}$. To obtain his Assumption 2(d), define the vector $\overline{S}_{it} = \left(S_{it}^{1}, ..., S_{it}^{L}\right)'$ and impose the following condition:

Condition 9 $E\left[\overline{S}_{it}\overline{S}'_{it} \mid C_t\right] - E\left[\overline{S}_{it} \mid C_t\right] E\left[\overline{S}_{it} \mid C_t\right]' > 0$ *almost surely.*

Condition 9 implies Assumption 2(d) in Andrews (2005) directly.

Next, define $\widehat{\beta} = \left(\widehat{\beta}_0, ..., \widehat{\beta}_L\right)$ and $\beta(C_t) = (\beta_0(C_t), ..., \beta_L(C_t))$. Theorem 4(a) in Andrews (2005) shows that, under his Assumptions 1-3,

$$\sqrt{N_t}\left(\widehat{\beta} - \beta(C_t)\right) \xrightarrow{d} V_C^{1/2} \times N(0, I). \tag{26}$$

where I is the identity matrix and

$$V_C = B_C^{-1}\Omega_C B_C^{-1} \tag{27}$$

where

$$\begin{aligned} B_C &= E\left[\left(\overline{S}_{it} - E\left[\overline{S}_{it} \mid C_t\right]\right)\left(\overline{S}_{it} - E\left[\overline{S}'_{it} \mid C_t\right]\right)' \mid C_t\right], \\ \Omega_C &= E\left[\xi_{it}\xi'_{it} \mid C_t\right] \end{aligned}$$

and

$$\begin{aligned} \xi_{it} &= \left(\overline{S}_{it} - E\left[\overline{S}_{it} \mid C_t\right]\right)\zeta_{it} \\ &\quad -E\left[\left(\overline{S}_{it} - E\left[\overline{S}_{it} \mid C_t\right]\right)\zeta_{it} \mid C_t\right] \\ &\quad -\left(\overline{S}_{it} - E\left[\overline{S}_{it} \mid C_t\right]\right)E\left[\zeta_{it} \mid C_t\right], \\ \zeta_{it} &= 1\left\{Y_{it} \leq \widetilde{y}\right\} - \Pr\left(Y_{it} \leq \widetilde{y} \mid S^1_{it}, ..., S^L_{it}, C_t\right). \end{aligned}$$

Note that ζ_{it} is the error in the linear model, and so $\Pr\left(\zeta_{it} \mid S^1_{it}, ..., S^L_{it}, C_t\right) = 0$. Next we turn to the proof of Proposition 1 in the main text.

Proof of Proposition 1. Condition 1 in the main text and Condition 9 imply the asymptotic distribution in (26) by applying Theorem 4(a) in Andrews (2005). From (26), it is clear that, for sufficiently large N_t,

$$\begin{aligned} V_C &= Var\left(\sqrt{N_t}\left(\widehat{\beta} - \beta(C_t)\right) \mid C_t\right) \\ &= N_t \times Var\left(\widehat{\beta} \mid C_t\right). \end{aligned}$$

and that V_C is an almost surely finite matrix.

To obtain

$$E\left(\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 \mid C_t\right) = e_t(s, C_t) \times N_t^{-1}, \tag{28}$$

where $e_t(s, C_t)$ is a $\sigma(C_t)$-measurable random variable that is almost surely finite, first note that, for any $s = 1, .., L$,

$$\begin{aligned} E\left(\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 \mid C_t\right) &= Var\left(\widehat{P}_t(s) \mid C_t\right) \\ &= Var\left(\widehat{\beta}_0 + \widehat{\beta}_s \mid C_t\right), \end{aligned}$$

because $\widehat{P}_t(s)$ is an unbiased estimator of $P_t(s, C_t)$. So,

$$\begin{aligned} & E\left(\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 \mid C_t\right) \\ = & Var\left(\widehat{\beta}_0 \mid C_t\right) + Var\left(\widehat{\beta}_s \mid C_t\right) + 2Cov\left(\widehat{\beta}_0, \widehat{\beta}_s \mid C_t\right) \\ = & \left[V_C(1,1) + V_C(s,s) + 2V_C(1,s)\right] \times \frac{1}{N_t} \end{aligned}$$

where $V_C(s, l)$ corresponds to the (s, l) element of the conditional covariance matrix V_C. The result follows by taking

$$\begin{aligned} e_t(0, C_t) &= V_C(1,1), \text{ and} \\ e_t(s, C_t) &= V_C(1,1) + V_C(s,s) + 2V_C(1,s), \end{aligned} \tag{29}$$

for any $s = 1, .., L$. ■

A.1.2 Continuous Case

In this subsection we employ the results of Souza-Rodrigues (2013) to show Proposition 2 in the main text. First, select a group t and let $\widehat{P}_t(s)$ be the Nadaraya-Watson kernel regression estimator

$$\widehat{P}_t(s) = \frac{\sum_{i=1}^{N_t} 1\{Y_{it} \leq \widetilde{y}\} K\left(\frac{S_{it}-s}{b}\right)}{\sum_{i=1}^{N_t} K\left(\frac{S_{it}-s}{b}\right)} \tag{30}$$

where $K(\cdot)$ is the kernel function, b is the bandwidth and N_t is the number of observations within group t. We want to show that $\widehat{P}_t(s)$ converges in probability (with rate) to the random limit $P_t(s, C_t) \equiv \Pr(Y_{it} \leq \widetilde{y} | S_{it} = s, C_t)$.[16]

We start by imposing existence of conditional density given the common shock. Then, we adapt the standard conditions used in the kernel literature to the present case. They are similar to the usual conditions, but they must hold for Q_t-almost all $c \in \mathcal{C}$.

[16] Following the notation presented in the Appendix A.3, the argument here holds for a given selected group $\tau \in \mathcal{T}$. In case the groups are selected randomly, then we put the argument conditional on the event $\{\tau_t(\omega) = \tau\}$, for any $t \geq 1$.

Condition 10 *There exists a conditional density of* (Y_{it}^*, S_{it}) *given* C_t*, denoted by* $f_t(y^*, s|c)$ *where* $y^* \in \mathcal{Y}^*$*,* $s \in \mathcal{S}$ *and* $c \in \mathcal{C}$*.*

Next we specify the restrictions on the kernel function and the bandwidth:

Condition 11 *Let* K *be the class of all Borel measurable nonnegative bounded real-valued functions* $K(\psi)$ *such that (i)* $\int K(\psi)d\psi = 1$*; (ii)* $|K(\psi)| \|\psi\|^k \to 0$ *as* $\|\psi\| \to \infty$*; (iii)* $\int K^2(\psi)d\psi < \infty$*; (iv)* $\sup_\psi |K(\psi)| < \infty$*; (v)* $\int \psi^2 K(\psi)d\psi < \infty$ *and (vi)* $\int \psi K(\psi)d\psi = 0$*.*

Condition 12 *(i)* $b \to 0$ *as* $N_t \to \infty$ *and (ii)* $N_t b \to \infty$ *as* $N_t \to \infty$*, for all* $t \geq 1$*.*

As usual, we need restrictions on the density of the regressors. In the present case, this translates into restrictions on the conditional density $f_t(s|c)$.

Condition 13 *For* Q_t*-almost all* $c \in \mathcal{C}$*, (i) the point* s *is in the interior of the support of* S *conditional on the event* $\{C_t = c\}$ *and (ii)* $f_t(s|c) \in [\upsilon, \infty)$*, for some finite* $\upsilon > 0$*.*

Condition 14 *For* Q_t*-almost all* $c \in \mathcal{C}$*, the conditional density* $f_t(s|c)$ *is continuous at any* s*.*

Condition 15 *For* Q_t*-almost all* $c \in \mathcal{C}$*, (i)* $f_t(s|c)$ *is twice continuously differentiable with respect to* s *in some neighborhood of* s *and (ii) the second-order derivatives of* $f_t(s|c)$ *with respect to* s *are bounded in this neighborhood.*

Condition 1 in the main text and Conditions 10-14 are sufficient to guarantee that the Nadaraya-Watson kernel density estimator

$$\widehat{f}_t(s) = \frac{1}{N_t b} \textstyle\sum_{i=1}^{N_t} K\left(\frac{S_{it} - s}{b}\right) \tag{31}$$

converges in probability to the random object $f_t(s|C_t)$ as $N_t \to \infty$. Conditions 13-15 are important to obtain its rate of convergence.

To obtain consistency with rate of $\widehat{P}_t(s)$ to $P_t(s, C_t)$ we impose the additional restriction:

Condition 16 *(i)* $U_t(\cdot)$ *is twice continuously differentiable, with bounded second derivatives, in some neighborhood of* s *and (ii) the conditional density of* Y_{it}^* *given* $(S_{it}, X_t, U_t(S_{it}))$*, denoted by* $f_t(y^*|s, x, u(s))$*, is twice continuously differentiable, with bounded second derivative, with respect to* $(s, u(\cdot))$*, for almost all* $(y^*, s, x, u(\cdot))$*.*

Condition 16 can be used to apply (Q_t-almost sure) Taylor expansions as is usually done in the literature. We need $P_t(s,c)$ to be twice continuously differentiable with respect to s, for Q_t-almost all c. Because

$$\begin{aligned} P_t(s,c) &= \int_{-\infty}^{D^{-1}(\widetilde{y})} f(y^*|s,c)\,dy^* \\ &= \int_{-\infty}^{D^{-1}(\widetilde{y})} f_t(y^*|s,x,u(s))\,dy^*, \end{aligned} \tag{32}$$

we need differentiability of $f_t(y^*|s,x,u(s))$ with respect to $(s,u(\cdot))$ as well as differentiability of $u(\cdot)$ with respect to s.[17]

Next, we prove Proposition 2 in the main text.

Proof of Proposition 2. The result follows directly from Proposition 4 in Souza-Rodrigues (2013). Here we present explicitly the quantities involved in the result. First, note that the conditional mean square error is given by

$$MSE\left(\widehat{P}_t(s) \mid C_t\right) = \left[E\left(\widehat{P}_t(s) \mid C_t\right) - P_t(s,C_t)\right]^2 + Var\left(\widehat{P}_t(s) \mid C_t\right)$$

From Lemma 4 in Souza-Rodrigues (2013), the bias term is given by

$$\begin{aligned} E\left(\widehat{P}_t(s) \mid C_t\right) - P_t(s,C_t) &= \frac{b^2}{2}\left(\int \psi^2 K(\psi)d\psi\right)\phi_t(s,C_t) \\ &\quad + O_p\left(\frac{1}{N_t b}\right) + o_p(b^2) \end{aligned}$$

where the function $\phi_t(s,c)$ is defined to be

$$\phi_t(s,c) = \left[f_t(s|c)\,\nabla_s^2 P_t(s,c) + 2\left[\nabla_s P_t(s,c)\right]\left[\nabla_s f_t(s|c)\right]\right]$$

and ∇_s^a denotes the a-th partial derivative with respect to s.

The variance term is

$$Var\left(\widehat{P}_t(s) \mid C_t\right) = \frac{1}{N_t b^{d_s}} \frac{\sigma_t^2(s,C_t)}{f_t(s|C_t)}\left(\int K^2(\psi)d\psi\right) + o_p\left(\frac{1}{N_t b^{d_s}}\right)$$

where $\sigma_t^2(s,c) \equiv Var\left[1\left\{Y_{it} \leq \widetilde{y}\right\} \mid S_{it} = s, C_t = c\right]$. The expressions for both the bias and variance come from a (Q_t-almost surely) Taylor expansion argument.

Note that (i) the term $\phi_t(s,c)$ is finite for Q_t-almost all $c \in \mathcal{C}$, by Conditions 15 and 16; (ii) both $\left(\int \psi^2 K(\psi)d\psi\right)$ and $\left(\int K^2(\psi)d\psi\right)$ are finite by Condition 11; and (iii) $f_t(s|c) \geq v > 0$, for

[17] Note that to obtain consistency of the kernel density $\widehat{f}_t(s)$ to the random object $f_t(s \mid C_t)$, we need assumptions on the conditional density $f_t(s \mid c) = f_t(s \mid x, u(\cdot), z)$. In addition, to obtain consistency of the kernel regression $\widehat{P}_t(s)$ to the Kolmogorov conditional expectation $P_t(s,C_t)$ we need assumptions on the conditional density $f_t(y^* \mid s, x, u(s))$.

Q_t-almost all $c \in \mathcal{C}$, by Condition 13. Therefore, we can rewrite the bias and variance terms as

$$E\left(\widehat{P}_t\left(s\right)\ \mid C_t\right) - P_t\left(s, C_t\right) = \frac{b^2}{2}\alpha_t\left(s, C_t\right) + O_p\left(\frac{1}{N_t b}\right) + o_p(b^2)$$

and

$$Var\left(\widehat{P}_t\left(s\right)\ \mid C_t\right) = \frac{1}{N_t b^{d_s}}\beta_t\left(s, C_t\right) + o_p\left(\frac{1}{N_t b^{d_s}}\right)$$

where the random variables $\alpha_t\left(s, C_t\right)$ and $\beta_t\left(s, C_t\right)$ are clearly defined from the context and both are Q_t-a.s. finite. Therefore, the conditional MSE is

$$\begin{aligned} MSE\left(\widehat{P}_t\left(s\right)\ \mid C_t\right) &= b^4\left[\frac{\alpha_t\left(s, C_t\right)}{2}\right]^2 + \frac{1}{N_t b^{d_s}}\beta_t\left(s, C_t\right) \\ &\quad + O_p\left(\frac{1}{N_t b}\right)^2 + o_p(b^4) + o_p\left(\frac{1}{N_t b^{d_s}}\right). \end{aligned} \tag{33}$$

By choosing the bandwidth $b \propto N_t^{-\frac{1}{4+d_s}}$, we conclude

$$MSE\left(\widehat{P}_t\left(s\right)\ \mid C_t\right) = N_t^{-\frac{4}{4+d_s}} \times \left[\left(\frac{\alpha_t\left(s, C_t\right)}{2}\right)^2 + \beta_t\left(s, C_t\right) + O_p\left(1\right)\right]. \tag{34}$$

The result follows by taking

$$e_t\left(s, C_t\right) = \left(\frac{\alpha_t\left(s, C_t\right)}{2}\right)^2 + \beta_t\left(s, C_t\right) + O_p\left(1\right). \tag{35}$$

■

A.1.3 Uniform Rate

Proof of Proposition 3. We want to show

$$E\left[\max_{1\le t\le T}\left|\widehat{P}_t\left(s\right) - P_t\left(s, C_t\right)\right|^2\right] \le const.\min\left\{\frac{T}{N^{2r}_{\min,t}}, 1\right\},$$

where $N_{\min,T} \equiv \min\left\{N_1, ..., N_T\right\}$. The expectation $E\left[\max_{1\le t\le T}\left|\widehat{P}_t\left(s\right) - P_t\left(s, C_t\right)\right|^2\right]$ is taken over the vector

$$\left\{\left\{Y_{it}, S_{it}\right\}_{i=1}^{N_t}, C_t\right\}_{t=1}^{T}$$

where $C_t = \left(X_t, U_t\left(\cdot\right), Z_t\right)$. First, note that

$$\begin{aligned} & E\left[\max_{1\le t\le T}\left|\widehat{P}_t\left(s\right) - P_t\left(s, C_t\right)\right|^2\right] \\ \le\ & T\max_{1\le t\le T} E\left[\left|\widehat{P}_t\left(s\right) - P_t\left(s, C_t\right)\right|^2\right] \\ =\ & T\max_{1\le t\le T} E\left[E\left[\left|\widehat{P}_t\left(s\right) - P_t\left(s, C_t\right)\right|^2 |C_t\right]\right] \end{aligned}$$

where the inequality follows by applying Lemma 2.2.2 in Van de Vaart and Wellner (1996); the expectation in the second line, $E\left[\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2\right]$, is taken over $\left(\{Y_{it}, S_{it}\}_{i=1}^{N_t}, C_t\right)$ because $\{Y_{it}, S_{it}, C_t : t \geq 1\}$ are independent across t; and the equality follows from the Law of Iterated Expectations. From Propositions 1 and 2, we have that

$$E\left[\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2 |C_t\right] = e(s, C_t) \times N_t^{-2r}.$$

where $r = 1/2$ if S_{it} is discrete, and $r = 2/(4 + d_s)$ if S_{it} is a d_s-vector of continuously distributed variables. Therefore,

$$\begin{aligned} & E\left[\max_{1\leq t\leq T}\left|\widehat{P}_t(s) - P_t(s, C_t)\right|^2\right] \\ \leq\ & T \max_{1\leq t\leq T} E\left[e(s, C_t) \times N_t^{-2r}\right] \\ \leq\ & \frac{T}{N_{\min,T}^{2r}} \times \max_{1\leq t\leq T} E\left[e(s, C_t)\right] \\ \leq\ & \frac{T}{N_{\min,T}^{2r}} \times \sup_{t\geq 1} E\left[e(s, C_t)\right] \\ \leq\ & const. \frac{T}{N_{\min,T}^{2r}} \end{aligned}$$

where the last inequality follows from (17), i.e. $\sup_{t\geq 1} E\left[e_t(s, C_t)\right] < \infty$. Because $E\left[\left|\widehat{P}_t(s) - P_t(s, C_t)\right|\right] \leq 1$, for any $t \leq T$, the result follows. ■

Remark 9 *Note from (35) that sufficient for* $\sup_{t\geq 1} E\left[e_t(s, C_t)\right] < \infty$ *(equation (17) in the main text), in case* S_{it} *is continuous, is to assume for all* $t \geq 1$ *and for* Q_t*-almost all* c*, (i)* $f_t(s|c) \leq K_0$ *and* $\nabla_s f_t(s|c) \leq K_1$*; (ii)* $\nabla_s P_t(s, c) \leq B_1$ *and* $\nabla_s^2 P_t(s, c) \leq B_2$*, for some finite constants* $K_0, K_1, B_1, B_2 > 0$*. Then, given Condition 11, and because* $\sigma_t^2(s, c) \leq 1$ *and* $f_t(s|c) \geq \upsilon > 0$ *by Condition 13, the term* $\sup_{t\geq 1} E\left[e_t(s, C_t)\right] < \infty$*.*

A.2 Second Step - Across Groups

Next we present the results for the second step.

A.2.1 Consistency

To simplify notation, let $P_t(s) \equiv P_t(s, C_t)$. Sometimes we omit both indices s and u from both $h_s(\cdot, u)$ and $\rho_u(\cdot)$ when it is clear enough from the context. Below we will make extensive use of the C_r-inequality: $(a + b)^r \leq c(a^r + b^r)$, where $c = 1$ if $r \leq 1$ and $c = 2^{r-1}$ if $r > 1$.

Define the unfeasible PSMD estimator by

$$\overline{m}(z,h) = p^{J_T}(z)'(P'P)^{-}\sum_{t=1}^{T} p^{J_T}(Z_t)\rho\left(P_t\left(s\right), X_t; h\right),$$

and the projection of $m\left(z,h\right)$ on $p^{J_T}(z)$ by

$$\widetilde{m}\left(z,h\right) = p^{J_T}(z)'(P'P)^{-}\sum_{t=1}^{T} p^{J_T}(Z_t)m\left(Z_t, h\right).$$

Let $N\left(\omega, \mathcal{F}, \|\cdot\|\right)$ be the covering number of a class of functions $\mathcal{F}$ under the norm $\|\cdot\|$ and $N_{[]}\left(\omega, \mathcal{F}, \|\cdot\|\right)$ be the corresponding bracketing number. For $j = 1, ..., J_T$, define the class of functions

$$\mathcal{O}_j = \left\{p_j\left(Z_t\right)\rho\left(P_t\left(s\right), X_t, h\right) : h \in \mathcal{H}_{K_T}\right\},$$

where the envelope of $\mathcal{O}_j$ is defined by $F_j\left(Z_t\right) = \overline{a}\left|p_j\left(Z_t\right)\right|$, for some finite $\overline{a} \geq 1$, because $\rho\left(P_t\left(s\right), X_t, h\right) \leq 1$ for all $h \in \mathcal{H}_{K(T)}$. Let

$$\Phi_{jT} = \int_0^1 \sqrt{1 + \log N_{[]}\left(\omega, \mathcal{O}_{jT}, \|\cdot\|_{L^2\left(f_{P,X,Z}\right)}\right)}d\omega.$$

We will need a bound for $\Phi_T \geq \max_{1\leq j\leq J_T}\Phi_{jT}$. By Remark C.3 in Chen and Pouzo (2012), it is possible to show that

$$\log N_{[]}\left(\omega, \mathcal{O}_{jT}, \|\cdot\|_{L^2\left(f_{P,X,Z}\right)}\right) \leq \log N\left(\omega, \mathcal{H}_{K_T}, \|\cdot\|_{L^2\left(f_{P,X,Z}\right)}\right),$$

and,

$$\begin{aligned}\log N\left(\omega, \mathcal{H}_{K_T}, \|\cdot\|_{L^2\left(f_{P,X,Z}\right)}\right) &\leq \log N_{[]}\left(\omega, \mathcal{H}_{K_T}, \|\cdot\|_{L^2\left(f_{P,X,Z}\right)}\right) \\ &\leq cons.\omega^{-d_x/\alpha},\end{aligned}$$

where the second inequality above follows from Theorem 1 and Corollaries 3.2 and 4.2 in Nickl and Pötscher (2007). So, $\Phi_T \leq cons.$, as required by Chen and Pouzo (2012).

Next we turn to the proof of Proposition 4 in the main text.

Proof of Proposition 4. The proof is based on Theorem 3.2 and Proposition 6.1 of Chen and Pouzo (2012). If $P_t(s)$ was observed in the data, the argument here would follow directly from their proof of Proposition 6.1, except for the necessary adjustments in the parameter space $\mathcal{H}$, the sieves space $\mathcal{H}_K$, the choice of the weighted sup-norm and the penalization term. The adjustments are necessary because we handle X_t with support on the real line (large support assumption), while they assumed $\mathcal{X}$ is compact. But the adjustments are straightforward. By inspecting the proofs of their Theorem 3.2 and Proposition 6.1, it is clear that we have imposed all the (adjusted) conditions they assumed, and so, if we had $P_t(s)$, instead of $\widehat{P}_t(s)$, consistency of $\widehat{h}_s(u)$ under the weighted sup-norm would follow directly. The presence of $\widehat{P}_t(s)$ changes their result in how the uniform convergence over $h \in \mathcal{H}_K$ of the criterion function to its populational version is obtained (i.e., their Assumption 3.3). More specifically, it is sufficient to prove that, for some finite constants $c_0, c > 0$,

$$c_0 E \left\| m\left(Z_t, h\right) \right\|_E^2 + O_p\left(\overline{\delta}_{m,T}^2\right) \leq \frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) \right\|_E^2 \leq cE \left\| m\left(Z_t, h\right) \right\|_E^2 + O_p\left(\overline{\delta}_{m,T}^2\right) \qquad (36)$$

uniformly over $\mathcal{H}_K$ for some $\overline{\delta}_{m,T}^2 = o(1)$.

To obtain the first inequality of (36) we use the fact that $(a-b)^2 + b^2 \geq \frac{1}{2}a^2$. This fact implies that uniformly over $h \in \mathcal{H}_K$,

$$\frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) \right\|_E^2 \geq \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T} \left\| \widetilde{m}\left(Z_t, h\right) \right\|_E^2 - \frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) - \widetilde{m}\left(Z_t, h\right) \right\|_E^2 .$$

Using $(a-b)^2 + b^2 \geq \frac{1}{2}a^2$ again

$$\frac{1}{T}\sum_{t=1}^{T} \left\| \widetilde{m}\left(Z_t, h\right) \right\|_E^2 \geq \frac{1}{2}\frac{1}{T}\sum_{t=1}^{T} \left\| \overline{m}\left(Z_t, h\right) \right\|_E^2 - \frac{1}{T}\sum_{t=1}^{T} \left\| \overline{m}\left(Z_t, h\right) - \widetilde{m}\left(Z_t, h\right) \right\|_E^2 .$$

By Lemma C.2(ii) of Chen and Pouzo (2012), we have that, uniformly over $h \in \mathcal{H}_K$,

$$\frac{1}{T}\sum_{t=1}^{T} \left\| \overline{m}\left(Z_t, h\right) \right\|_E^2 \geq c_0 E \left\| m\left(Z_t, h\right) \right\|_E^2 - O_p\left(\frac{J_T}{T}\Phi_T + b_{mT}^2\right), \qquad (37)$$

where $\frac{J_T}{T}\Phi_T = o(1)$. So, $\left(\frac{J_T}{T}\Phi_T + b_{mT}^2\right) \to 0$ as $T \to \infty$.

In addition, by Lemma C.1(ii) of Chen and Pouzo (2012),

$$\sup_{h \in \mathcal{H}_K} \frac{1}{T}\sum_{t=1}^{T} \left\| \overline{m}\left(Z_t, h\right) - \widetilde{m}\left(Z_t, h\right) \right\|_E^2 = O_p\left(\frac{J_T}{T}\Phi_T\right) = o_p(1). \qquad (38)$$

Therefore, uniformly over $h \in \mathcal{H}_K$,

$$\frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) \right\|_E^2 \geq c_0 E \left\| m\left(Z_t, h\right) \right\|_E^2 - O_p\left(\frac{J_T}{T}\Phi_T + b_{mT}^2\right) - \frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) - \widetilde{m}\left(Z_t, h\right) \right\|_E^2 .$$

Similarly, to obtain the second inequality of (36) we use the fact that $(a-b)^2 \leq 2a^2 + 2b^2$. So, uniformly over $h \in \mathcal{H}_K$,

$$\frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) \right\|_E^2 \leq 2\frac{1}{T}\sum_{t=1}^{T} \left\| \widetilde{m}\left(Z_t, h\right) \right\|_E^2 + 2\frac{1}{T}\sum_{t=1}^{T} \left\| \widehat{m}\left(Z_t, h\right) - \widetilde{m}\left(Z_t, h\right) \right\|_E^2 .$$

Using $(a-b)^2 \leq 2a^2 + 2b^2$ again

$$\frac{1}{T}\sum_{t=1}^{T} \left\| \widetilde{m}\left(Z_t, h\right) \right\|_E^2 \leq 2\frac{1}{T}\sum_{t=1}^{T} \left\| \overline{m}\left(Z_t, h\right) \right\|_E^2 + 2\frac{1}{T}\sum_{t=1}^{T} \left\| \overline{m}\left(Z_t, h\right) - \widetilde{m}\left(Z_t, h\right) \right\|_E^2 .$$

By Lemmas C.1(ii) and C.2(ii) of Chen and Pouzo (2012), we obtain uniformly over $h \in \mathcal{H}_K$

$$\frac{1}{T}\sum_{t=1}^{T} \|\widehat{m}(Z_t,h)\|_E^2 \leq cE\|m(Z_t,h)\|_E^2 + O_p\left(\frac{J_T}{T}\Phi_T + b_{mT}^2\right) + \frac{1}{T}\sum_{t=1}^{T}\|\widehat{m}(Z_t,h) - \widetilde{m}(Z_t,h)\|_E^2 .$$

Therefore, to obtain (36), we need to show

$$\sup_{h\in\mathcal{H}_K} T^{-1}\sum_{t=1}^{T}\|\widehat{m}(Z_t,h) - \widetilde{m}(Z_t,h)\|_E^2 = O_p(r_{TN}) \tag{39}$$

for some $r_{TN} = o(1)$ where

$$\widehat{m}(z,h) - \widetilde{m}(z,h) = p^{J_T}(z)'(P'P)^{-}\sum_{t=1}^{T} p^{J_T}(Z_t)\left[\rho\left(\widehat{P}_t(s), X_t; h\right) - m(Z_t,h)\right].$$

Define the random element

$$\eta_t = \left((Y_{it}, S_{it})_{i=1}^{N_t}, C_t\right),$$

where $C_t = (X_t, U_t(\cdot), Z_t)$. Recall that $\widehat{P}_t(s)$ is a statistic that depends on the sample $(Y_{it}, S_{it})_{i=1}^{N_t}$ and that $P_t(s, C_t)$ is measurable with respect to $\sigma(C_t)$. Define the residual function difference $\widehat{\rho}(\eta_t, h) = \left[\rho\left(\widehat{P}_t(s), X_t, Z_t\right) - m(Z_t,h)\right]$ and the T-vector $\widehat{\rho}(h) = [\widehat{\rho}(\eta_1,h), ..., \widehat{\rho}(\eta_T,h)]$. Now notice that

$$\begin{aligned}
& \sup_{h\in\mathcal{H}_K} T^{-1}\sum_{t=1}^{T}\|\widehat{m}(Z_t,h) - \widetilde{m}(Z_t,h)\|_E^2 \\
= & \sup_{h\in\mathcal{H}_K} T^{-1}\sum_{t=1}^{T} Tr\left\{p^{J_T}(Z_t)'(P'P)^{-}P'\widehat{\rho}(h)\widehat{\rho}(h)'P(P'P)^{-}p^{J_T}(Z_t)\right\} \\
= & \sup_{h\in\mathcal{H}_K} T^{-1}Tr\left\{\widehat{\rho}(h)'P(P'P)^{-}P'\widehat{\rho}(h)\right\} \\
\leq & \frac{1}{\lambda_{\min}(P'P/T)} \times \sup_{h\in\mathcal{H}_K}\frac{1}{T^2}Tr\left\{\widehat{\rho}(h)'PP'\widehat{\rho}(h)\right\} \\
= & \frac{1}{\lambda_{\min}(P'P/T)} \times \sup_{h\in\mathcal{H}_K}\left[\sum_{j=1}^{J_T}\left(\frac{1}{T}\sum_{t=1}^{T} p_j(Z_t)\widehat{\rho}(\eta_t,h)\right)^2\right].
\end{aligned} \tag{40}$$

Because $(\lambda_{\min}(P'P/T))^{-1} = O_p(1)$, by Condition 7(ii), we can focus on the second term of (40) to determine r_{TN} in (39). Let $M > 0$, then, by Markov inequality,

$$\begin{aligned}
& \Pr\left(\sup_{h\in\mathcal{H}_K}\left[\sum_{j=1}^{J_T}\left(\frac{1}{T}\sum_{t=1}^{T} p_j(Z_t)\widehat{\rho}(\eta_t,h)\right)^2\right] > r_{TN}M\right) \\
\leq & \frac{1}{r_{TN}M}E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left[\sum_{j=1}^{J_T}\left(\frac{1}{T}\sum_{t=1}^{T} p_j(Z_t)\widehat{\rho}(\eta_t,h)\right)^2\right]\right] \\
\leq & \frac{1}{r_{TN}M}\frac{J_T}{T}\max_{1\leq j\leq J_T} E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} p_j(Z_t)\widehat{\rho}(\eta_t,h)\right)^2\right]
\end{aligned} \tag{41}$$

where $\eta^T = (\eta_1, ..., \eta_T)$ and $E_{\eta^T}[\cdot]$ is the expectation taken over η^T. Next we centralize the process for each $j = 1, ..., J_T$ and use the inequality $(a-b)^2 \leq 2a^2 + 2b^2$ to obtain

$$\begin{aligned}
& E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} p_j(Z_t)\,\widehat{\rho}(\eta_t, h)\right)^2\right] \\
\leq \quad & 2E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} p_j(Z_t)\,\widehat{\rho}(\eta_t, h) - E_{\eta_t}\left[p_j(Z_t)\,\widehat{\rho}(\eta_t, h)\right]\right)^2\right] \\
& +2\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}E_{\eta^T}\left(\sum_{t=1}^{T}\left[p_j(Z_t)\,\widehat{\rho}(\eta_t, h)\right]\right)\right)^2\right].
\end{aligned} \tag{42}$$

It is sufficient therefore to bound (42). We start with the second term of the RHS of (42).

(I) **Bound The Second Term in (42).**

Denote $\widehat{A}_T(s) = \max_{t\leq T}\left|\widehat{P}_t(s) - P_t(s)\right|$. Note that

$$\begin{aligned}
\rho\left(\widehat{P}_t(s), X_t; h\right) &= 1\left\{\widehat{P}_t(s) \leq h(X_t, u)\right\} \\
&= 1\left\{P_t(s) - h(X_t, u) \leq P_t(s) - \widehat{P}_t(s)\right\} \\
&\leq 1\left\{P_t(s) - h(X_t, u) \leq \max_{t\leq T}\left|\widehat{P}_t(s) - P_t(s)\right|\right\}.
\end{aligned} \tag{43}$$

Then

$$\begin{aligned}
E_{\eta^T}\left[\sum_{t=1}^{T} p_j(Z_t)\,\widehat{\rho}(\eta_t, h)\right] &= E_{\eta^T}\left[\sum_{t=1}^{T} p_j(Z_t)\left[\rho\left(\widehat{P}_t(s), X_t; h\right) - m(Z_t, h)\right]\right] \\
&\leq E_{\eta^T}\left[\sum_{t=1}^{T} p_j(Z_t)\left(1\left\{P_t(s) - h(X_t, u) \leq \widehat{A}_T(s)\right\} - m(Z_t, h)\right)\right].
\end{aligned}$$

Moreover, using the law of iterated expectations twice,

$$\begin{aligned}
& E_{\eta^T}\left[\sum_{t=1}^{T} p_j(Z_t)\left(1\left\{P_t(s) - h(X_t, u) \leq \widehat{A}_T(s)\right\} - m(Z_t, h)\right)\right] \\
= \quad & E_{\eta^T}\left[\sum_{t=1}^{T} p_j(Z_t)\left(E_{\eta^T}\left(1\left\{P_t(s) \leq h(X_t, u) + \widehat{A}_T(s)\right\}|Z^T\right) - m(Z_t, h)\right)\right] \\
= \quad & E_{\eta^T}\left[\sum_{t=1}^{T} p_j(Z_t)\left(E_{\eta^T}\left(F_{P_{s/X,Z}}\left(h(X_t, u) + \widehat{A}_T(s)|X_t, Z_t\right)|Z^T\right) - m(Z_t, h)\right)\right],
\end{aligned}$$

where $F_{P_s|X,Z}$ is the conditional distribution function of $P_t(s)$ across t, $Z^T = (Z_1, ..., Z_T)$ and

$E_{\eta^T}\left[\cdot|Z^T\right]$ is the conditional expectation of η^T given Z^T. By the mean-value theorem,

$$\begin{aligned}
& E_{\eta^T}\left(F_{P_s/X,Z}\left(h\left(X_t,u\right)+\widehat{A}_T(s)|X_t,Z_t\right)|Z^T\right) \\
= \;& E_{\eta^T}\left(F_{P_s/X,Z}\left(h\left(X_t,u\right)|X_t,Z_t\right)|Z^T\right) \\
& +E_{\eta^T}\left(\left\{\int_0^1 f_{P_s/X,Z}\left(h\left(X_t,u\right)+t\widehat{A}_T(s)|X_t,Z_t\right)dt\right\}\times\left[\widehat{A}_T\left(s\right)\right]|Z^T\right) \\
\leq \;& m\left(Z_t,h\right)+(1-u)+E_{\eta^T}\left(\sup_{t\in[0,1]} f_{P_s/X,Z}\left(h\left(X_t,u\right)+t\widehat{A}_T(s)|X_t,Z_t\right)\times\left[\widehat{A}_T\left(s\right)\right]|Z^T\right) \\
\leq \;& m\left(Z_t,h\right)+1+KE_{\eta^T}\left(\widehat{A}_T\left(s\right)|Z^T\right)
\end{aligned}$$

where the first inequality uses the definition of $m\left(Z_t,h\right)$ and the fact that Z_t is i.i.d.; and the third inequality uses Condition 8(i). And so,

$$\begin{aligned}
& E_{\eta^T}\left[\sum_{t=1}^T p_j\left(Z_t\right)\left(E_{\eta^T}\left(F_{P_s/X,Z}\left(h\left(X_t,u\right)+\widehat{A}_T(s)|X_t,Z_t\right)|Z^T\right)-m\left(Z_t,h\right)\right)\right] \\
\leq \;& KE_{\eta^T}\left[\sum_{t=1}^T p_j\left(Z_t\right)\left(1+E_{\eta^T}\left(\widehat{A}_T\left(s\right)|Z^T\right)\right)\right].
\end{aligned}$$

Therefore, the second term on the RHS of the inequality (42) is such that

$$\begin{aligned}
& \sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}E_{\eta^T}\left(\sum_{t=1}^T\left[p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)\right]\right)\right)^2 \\
\leq \;& K^2\left(\frac{1}{\sqrt{T}}E_{\eta^T}\left[\sum_{t=1}^T p_j\left(Z_t\right)\left(1+E_{\eta^T}\left(\widehat{A}_T\left(s\right)|Z^T\right)\right)\right]\right)^2 \\
\leq \;& const.\frac{1}{T}E_{\eta^T}\left[\sum_{t=1}^T\left(p_j\left(Z_t\right)\right)^2\right] \\
& +const.\frac{1}{T}E_{\eta^T}\left[\sum_{t=1}^T\left(p_j\left(Z_t\right)\right)^2 E_{\eta^T}\left(\left(\widehat{A}_T\left(s\right)\right)^2|Z^T\right)\right]
\end{aligned}$$

where the second inequality uses both Jensen's inequality and $(a+b)^2\leq 2a^2+2b^2$.

By Conditions 2(i) and 7(ii),

$$\frac{1}{T}E_{\eta^T}\left[\sum_{t=1}^T\left(p_j\left(Z_t\right)\right)^2\right]\leq const<\infty.$$

Remember that $\xi_T\equiv\sup_{z\in\mathcal{Z}}\left\|p^{J_T}\left(z\right)\right\|_E$. Then

$$\begin{aligned}
& \frac{1}{T}E_{\eta^T}\left[\sum_{t=1}^T\left(p_j\left(Z_t\right)\right)^2 E_{\eta^T}\left(\left(\widehat{A}_T\left(s\right)\right)^2|Z^T\right)\right] \\
\leq \;& \frac{1}{T}E_{\eta^T}\left[T\left(\sup_{z\in\mathcal{Z}}\left\|p^{J_T}\left(z\right)\right\|_E^2\right)\times E_{\eta^T}\left(\left(\widehat{A}_T\left(s\right)\right)^2|Z^T\right)\right] \\
\leq \;& \xi_T^2\times E_{\eta^T}\left[\left(\widehat{A}_T\left(s\right)\right)^2\right] \\
\leq \;& \xi_T^2\times E_{\eta^T}\left[\max_{t\leq T}\left|\widehat{P}_t\left(s\right)-P_t\left(s\right)\right|^2\right]
\end{aligned}$$

where the second inequality uses the law of iterated expectations. In short, we obtained

$$\begin{aligned} & \sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}E_{\eta^T}\left(\sum_{t=1}^{T}\left[p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)\right]\right)\right)^2 \\ \leq\ & const\left(1+\xi_T^2\times E_{\eta^T}\left[\max_{t\leq T}\left|\widehat{P}_t\left(s\right)-P_t\left(s\right)\right|^2\right]\right). \qquad (44) \end{aligned}$$

(II) **Bound The First Term in (42).**

Next, we bound the term

$$E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)-E_{\eta_t}\left[p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)\right]\right)^2\right].$$

To simplify notation, define $\varepsilon_j\left(\eta_t,h\right)=p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)$ and

$$\overline{\varepsilon}_j\left(\eta_t,h\right)=\varepsilon_j\left(\eta_t,h\right)-E_{\eta_t}\left[\varepsilon_j\left(\eta_t,h\right)\right].$$

Note that the envelope of $\varepsilon_j\left(\eta_t,h\right)$ is the same as the envelope of $\mathcal{O}_j$, $F_j\left(Z_t\right)$. We have that

$$\begin{aligned} & E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\overline{\varepsilon}_j\left(\eta_t,h\right)\right)^2\right] \\ \leq\ & E_{\eta^T}\left[\left(\sup_{h\in\mathcal{H}_K}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\overline{\varepsilon}_j\left(\eta_t,h\right)\right)^2\right] \\ \leq\ & \left(E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\overline{\varepsilon}_j\left(\eta_t,h\right)\right|\right]+\sqrt{E_{\eta^T}\left[F_j\left(Z_t\right)^2\right]}\right)^2 \\ \leq\ & 2\left(E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\overline{\varepsilon}_j\left(\eta_t,h\right)\right|\right]\right)^2+2E_{\eta^T}\left[F_j\left(Z_t\right)^2\right], \qquad (45) \end{aligned}$$

where the second inequality follows from Theorem 2.14.5 in Van de Vaart and Wellner (1996); and the third inequality from the C_r-inequality. We know that $E_{\eta^T}\left[F_j\left(Z_t\right)^2\right]\leq const$ by Condition 7(ii). So we need to bound the first term of the RHS of (45).

To simplify notation, let $F_j\left(Z_t\right)=F_j$. Define the norm $\|.\|_{2,T}$ to be

$$\|g\|_{2,T}\equiv\left(\frac{1}{T}\sum_{t=1}^{T}E_{\eta_t}\left[g\left(\eta_t\right)\right]^2\right)^{1/2},$$

for some squared-integrable g, and

$$\begin{aligned} \mathcal{E}_{jT} &= \left\{\varepsilon_j\left(\eta_t,h\right):h\in\mathcal{H}_K\right\} \\ &= \left\{p_j\left(Z_t\right)\left[\rho\left(\widehat{P}_t\left(s\right),X_t;h\right)-m\left(Z_t,h\right)\right]:h\in\mathcal{H}_K\right\}. \end{aligned}$$

By Theorem 6 and Corollary 7 in Pollard (2002), the empirical process for independent (but not identically distributed) η_t is such that

$$\begin{aligned}
& E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\overline{\varepsilon}_j\left(\eta_t,h\right)\right|\right] \\
\leq\; & cE\left[F_j^2\right]\times\left(\int_0^1\sqrt{\log 2N_{[]}\left(\frac{E\left[F_j^2\right]}{2}\omega,\mathcal{E}_{jT},\left\|\cdot\right\|_{2,T}\right)}d\omega\right)
\end{aligned}$$

where c is some finite constant.[18]

We need to bound the bracketing number of $\mathcal{E}_{jT}$ under the norm $\left\|\cdot\right\|_{2,T}$. From Lemma 1 below, we have that

$$\int_0^1\sqrt{\log 2N_{[]}\left(\frac{E\left[F_j^2\right]}{2}\omega,\mathcal{E}_{jT},\left\|\cdot\right\|_{2,T}\right)}d\omega\leq const.$$

which implies

$$\begin{aligned}
& E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\overline{\varepsilon}_j\left(\eta_t,h\right)\right)^2\right] \\
\lesssim\; & \left(E\left[F_j^2\right]\right)^2\times\left(\int_0^1\sqrt{\log 2N_{[]}\left(\frac{E\left[F_j^2\right]}{2}\omega,\mathcal{E}_{jT},\left\|\cdot\right\|_{2,T}\right)}d\omega\right)^2+2E\left[F_j^2\right] \\
\leq\; & const. \qquad (46)
\end{aligned}$$

(III) **Collect Results from (I) and (II).**

We collect the results to bound the inequality (41). From inequalities (42), (44) and (46), we obtain

$$\begin{aligned}
& \Pr\left(\sup_{h\in\mathcal{H}_K}\left[\sum_{j=1}^{J_T}\left(\frac{1}{T}\sum_{t=1}^{T}p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)\right)^2\right]>r_{TN}M\right) \\
\leq\; & \frac{1}{r_{TN}M}\frac{J_T}{T}\max_{1\leq j\leq J_T}E_{\eta^T}\left[\sup_{h\in\mathcal{H}_K}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}p_j\left(Z_t\right)\widehat{\rho}\left(\eta_t,h\right)\right)^2\right] \\
\lesssim\; & \frac{1}{r_{TN}M}\frac{J_T}{T}\max_{1\leq j\leq J_T}\left\{\left(1+\xi_T^2\times E_{\eta^T}\left[\max_{t\leq T}\left|\widehat{P}_t\left(s\right)-P_t\left(s\right)\right|^2\right]\right)+const.\right\} \\
\lesssim\; & \frac{1}{r_{TN}M}\frac{J_T}{T}\left[1+\xi_T^2\min\left\{\frac{T}{N_{\min,T}^{2r}},1\right\}\right],
\end{aligned}$$

[18] The results in Pollard (2002) follow by replacing his norm $\left\|g\right\|_2=\left(\sum_{t=1}^{T}E_{\eta_t}\left[g\left(\eta_t\right)\right]^2\right)^{1/2}$ by the norm $\left\|g\right\|_{2,T}\equiv\left(\frac{1}{T}\sum_{t=1}^{T}E_{\eta_t}\left[g\left(\eta_t\right)\right]^2\right)^{1/2}$; and by replacing the inequality $\sup_x\left|g\left(x\right)\right|\leq\beta$ in his Lemma 3 by the inequality $\sup_x\left|g\left(x\right)\right|\leq\beta\sqrt{T}$.

where the last inequality comes from Proposition 3.

If $\xi_T^2 \asymp J_T$, and $K_T \asymp J_T$, then

$$\begin{aligned} &\frac{J_T}{T}\left[1+\xi_T^2 \min\left\{\frac{T}{N_{\min,T}^{2r}},1\right\}\right] \\ \lesssim &\left[\frac{K_T}{T}+\frac{K_T^2}{T}\min\left\{\frac{T}{N_{\min,T}^{2r}},1\right\}\right] \end{aligned}$$

and (23) implies that the both terms on the RHS above converge to zero as $(T, N_{\min,T}) \to \infty$. We can take

$$r_{TN} = \left[\frac{K_T}{T}+\frac{K_T^2}{T}\min\left\{\frac{T}{N_{\min,T}^{2r}},1\right\}\right]$$

and $r_{TN} \to 0$ as $(T, N_{\min,T}) \to \infty$ by (23). We obtain therefore

$$\sup_{h\in\mathcal{H}_K} \frac{1}{T}\sum_{t=1}^{T} \left\|\widehat{m}\left(Z_t,h\right)-\widetilde{m}\left(Z_t,h\right)\right\|_E^2 = O_p\left(r_{TN}\right) = o_p\left(1\right).$$

■

Next, we prove the following Lemma:

Lemma 1 *Define the norm* $\|g\|_{2,T} = \left(\frac{1}{T}\sum\limits_{t=1}^{T} E_{\eta_t}\left|g\left(\eta_t\right)\right|^2\right)^{1/2}$*, and the set of functions*

$$\mathcal{E}_{jT} = \left\{\varepsilon_j\left(\cdot,h\right) : h \in \mathcal{H}_{K(T)}\right\},$$

where $\varepsilon_j\left(\eta_t,h\right) = p_j\left(Z_t\right)\left[\rho\left(\widehat{P}_t\left(s\right),X_t;h\right)-m\left(Z_t,h\right)\right]$*. Let the conditions of Proposition 4 hold. Then*

$$\int_0^1 \sqrt{\log 2N_{[]}\left(\frac{E\left[F_j^2\right]}{2}\omega, \mathcal{E}_{jT}, \|\cdot\|_{2,T}\right)}d\omega \leq const. < \infty$$

Proof. We need to bound the bracketing number of $\mathcal{E}_{jT}$ under the norm $\|\cdot\|_{2,T}$. To do so, we split the function $\varepsilon_j\left(\eta_t,h\right)$ in two parts, $p_j\left(Z_t\right)\rho\left(\widehat{P}_t\left(s\right),X_t;h\right)$ and $p_j\left(Z_t\right)m\left(Z_t,h\right)$, and obtain the bracketing numbers for each class of functions. Then we bound each bracketing number by the covering number of $\mathcal{H}_K$. Finally, we combine them to obtain the desired result.

Define

$$\varrho_j = \left\{p_j\left(Z_t\right)\rho\left(\widehat{P}_t\left(s\right),X_t;h\right) : h \in \mathcal{H}_{K(T)}\right\}$$

and

$$\mathcal{M}_j = \left\{p_j\left(Z_t\right)m\left(Z_t,h\right) : h \in \mathcal{H}_{K(T)}\right\}.$$

Then, $\mathcal{E}_{jT} = \varrho_j \oplus \mathcal{M}_j$ and, by Theorem 6 of Andrews (1994),

$$\log N_{[]}\left(\omega, \mathcal{E}_{jT}, \|\cdot\|_{2,T}\right) \leq \log N_{[]}\left(\frac{\omega}{2}, \varrho_j, \|\cdot\|_{2,T}\right) + \log N_{[]}\left(\frac{\omega}{2}, \mathcal{M}_j, \|\cdot\|_{2,T}\right). \tag{47}$$

So,

$$\begin{aligned}
&\int_0^1 \sqrt{\log 2N_{[]}\left(\frac{E\left[F_j^2\right]}{2}\omega, \mathcal{E}_{jT}, \|\cdot\|_{2,T}\right)}\, d\omega \\
\lesssim\ & c_0 + \int_0^1 \sqrt{\log N_{[]}\left(\frac{E\left[F_j^2\right]}{2}\omega, \mathcal{E}_{jT}, \|\cdot\|_{2,T}\right)}\, d\omega \\
\leq\ & c_0 + \int_0^1 \sqrt{\log N_{[]}\left(\frac{E\left[F_j^2\right]}{4}\omega, \varrho_j, \|\cdot\|_{2,T}\right)}\, d\omega \\
& + \int_0^1 \sqrt{\log N_{[]}\left(\frac{E\left[F_j^2\right]}{4}\omega, \mathcal{M}_j, \|\cdot\|_{2,T}\right)}\, d\omega
\end{aligned}$$

where the first inequality follows from the C_r-inequality and the last inequality, from (47).

(A) Bound the Covering Number for $\mathcal{H}_K$.

Recall that $\|h\|_{\infty,\omega} = \sup_{x\in\mathcal{X}} |\omega(x) h(x)|$, with the weight function $\omega(x) = \left(1 + \|x\|_E^2\right)^{-\mu/2}$, for some $\mu > 0$. Define another weighted norm $\|h\|_{\infty,\varpi} = \sup_{x\in\mathcal{X}} |\varpi(x) h(x)|$, with

$$\varpi(x) = \left(1 + \|x\|_E^2\right)^{-(\mu+\kappa)/2},$$

with some $\kappa > 0$ and $\kappa \neq \alpha$ if $\mathcal{H}$ is the weighted Hölder space; and some $\kappa > \alpha - \frac{d_x}{2} > 0$ if $\mathcal{H}$ is the weighted Sobolev space. Because (i) $\mathcal{H}_K$ is a bounded subset of $\mathcal{H}$ under the weighted norm (Condition 5), (ii) and $E\left[\left(1 + \|X_t\|_E^2\right)^{2(\mu+\kappa)}\right] < \infty$ (Condition 8(ii)), we can apply Theorem 1 and Corollaries 3.2 and 4.2 in Nickl and Pötscher (2007) [see their equation (3), p.184].[19] The covering number of $\mathcal{H}_K$ under the norm $\|\cdot\|_{\infty,\varpi}$ is therefore

$$\log N\left(\omega, \mathcal{H}_K, \|\cdot\|_{\infty,\varpi}\right) \leq const.\omega^{-\frac{d_x}{\alpha}}.$$

(B) Bound the Bracketing Number for $\mathcal{M}_j$.

[19] Using Nickl and Pötscher's (2007) notation, the result here follows by Theorem 1 with (a) $p = q = \infty$, $r = 1$, and $\gamma \neq s > 0$, if $\mathcal{H}$ is the weighted Hölder (Corollary 3.2); and (b) $p = 2, q = \infty$, $r = 1$, and $\gamma > s - d/2 > 0$, if $\mathcal{H}$ is the weighted Sobolev (Corollary 4.2).

First note that $p_j(Z_t)\,m(Z_t,h)$ is $\|\cdot\|_{2,T}$-Lipschitz in $\left(\mathcal{H},\|\cdot\|_{\infty,\varpi}\right)$.

$$\begin{aligned}
& \left|p_j(z)\left(m(z,h)-m\left(z,h'\right)\right)\right| \\
\leq\ & |p_j(z)|\left|m(z,h)-m\left(z,h'\right)\right| \\
=\ & |p_j(z)|\left|F_{P_s|X,Z}(h(x)\,|x,z)-F_{P_s|X,Z}\left(h'(x)\,|x,z\right)\right| \\
\leq\ & |p_j(z)|\left|\left\{\int_0^1 f_{P_s/X,Z}\left(h(x)+th'(x)\,|X_t,Z_t\right)dt\right\}\right|\times\left|h(x)-h'(x)\right| \\
\leq\ & F_j(z)\times K\times\left|h(x)-h'(x)\right| \\
\leq\ & F_j(z)\times K\times\left|\frac{1}{\varpi(x)}\right|\times\left\|h-h'\right\|_{\infty,\varpi}
\end{aligned}$$

where the second inequality comes from the mean-value theorem and the third inequality comes from the definition of F_j and from Condition 8(i). So,

$$\begin{aligned}
& \left\|p_j(Z_t)\left(m(Z_t,h)-m\left(Z_t,h'\right)\right)\right\|_{2,T} \\
=\ & \left(\frac{1}{T}\sum_{t=1}^{T}E_{\eta_t}\left[p_j(Z_t)\left(m(Z_t,h)-m\left(Z_t,h'\right)\right)\right]^2\right)^{1/2} \\
\leq\ & \left(\frac{1}{T}\sum_{t=1}^{T}E_{\eta_t}\left[F_j(Z_t)\times K\times\left|\frac{1}{\varpi(X_t)}\right|\times\left\|h-h'\right\|_{\infty,\varpi}\right]^2\right)^{1/2} \\
=\ & \left(K^2E_{\eta_t}\left[F_j(Z_t)^2\left|\frac{1}{\varpi(X_t)}\right|^2\right]\times\left\|h-h'\right\|_{\infty,\varpi}^2\right)^{1/2} \\
\leq\ & \left(K^2\left(E_{\eta_t}\left[F_j(Z_t)^4\right]\right)^{1/2}\times\left(E_{\eta_t}\left[\left(1+\|X\|_E^2\right)^{2(\mu+\kappa)}\right]\right)^{1/2}\times\left\|h-h'\right\|_{\infty,\varpi}^2\right)^{1/2} \\
\leq\ & c\left\|h-h'\right\|_{\infty,\varpi},
\end{aligned}$$

where the third inequality follows from Cauchy-Schwartz inequality; and the last inequality from Conditions 7 and 8. By Theorem 2.7.11 in Van de Vaart and Wellner (1996),

$$\begin{aligned}
\log N_{[]}\left(\omega,\mathcal{M}_j,\|\cdot\|_{2,T}\right) &\leq \log N\left(\frac{\omega}{2c},\mathcal{H}_K,\|\cdot\|_{\infty,\varpi}\right) \\
&\leq \mathit{const.}\left(\frac{\omega}{2c}\right)^{-\frac{d_x}{\alpha}} \qquad (48)
\end{aligned}$$

Therefore

$$
\begin{aligned}
& \int_0^1 \sqrt{\log N_{[]}\left(\frac{E\left[F_j^2\right]}{4}\omega, \mathcal{M}_j, \left\|\cdot\right\|_{2,T}\right)} d\omega \\
\leq \quad & \int_0^1 \sqrt{\log N\left(\frac{E\left[F_j^2\right]}{8c}\omega, \mathcal{H}_K, \left\|\cdot\right\|_{\infty,\varpi}\right)} d\omega \\
\leq \quad & \int_0^1 const. \left(\frac{E\left[F_j^2\right]}{8c}\omega\right)^{-\frac{d_x}{2\alpha}} d\omega \\
\leq \quad & const.
\end{aligned}
$$

(C) **Bound the Bracketing Number for ϱ_j.**

Next, we consider the bracketing number for $\varrho_j = \left\{p_j\left(Z_t\right)\rho\left(\widehat{P}_t\left(s\right), X_t; h\right) : h \in \mathcal{H}_{K(T)}\right\}$, where $\rho\left(\widehat{P}_t\left(s\right), X_t; h\right) = 1\left\{\widehat{P}_t\left(s\right) \leq h\left(X_t\right)\right\}$.

The argument here follows the argument of Corollary 2.7.3 in Van de Vaart and Wellner (1996). Let $h_1, ..., h_n$ be the centers of $\left\|\cdot\right\|_{\infty,\varpi}$-balls of radius δ that cover $\mathcal{H}_K$. I.e., n is the δ-covering number of $\left(\mathcal{H}_K, \left\|\cdot\right\|_{\infty,\varpi}\right)$. Define for $l = 1, ...n$, the brackets

$$
\begin{aligned}
A_l &= 1\left\{\widehat{P}_t\left(s\right) \leq h\left(X_t\right) - \delta\right\} \\
B_l &= 1\left\{\widehat{P}_t\left(s\right) \leq h\left(X_t\right) + \delta\right\}.
\end{aligned}
$$

Then $\left[A_l, B_l\right]$, for $l = 1, ...n$, are brackets that cover ϱ_j. The $\left\|\cdot\right\|_{2,T}$-size of the brackets is, for $l = 1, ...n$,

$$
\begin{aligned}
& \left\|A_l \triangle B_l\right\|_{2,T} \\
= \quad & \left(\frac{1}{T}\sum_{t=1}^{T} E_{\eta_t}\left[A_l \triangle B_l\right]^2\right)^{1/2} \\
= \quad & \left(\frac{1}{T}\sum_{t=1}^{T} E_{\eta_t}\left[p_j\left(Z_t\right)^2 1\left\{h_l\left(X_t\right) - \delta \leq \widehat{P}_t\left(s\right) \leq h_l\left(X_t\right) + \delta\right\}\right]\right)^{1/2} \\
\leq \quad & \left(\frac{1}{T}\sum_{t=1}^{T} \left(E_{\eta_t}\left[F_j\left(Z_t\right)^4\right]\right)^{1/2} \times \left(E_{\eta_t}\left[1\left\{h_l\left(X_t\right) - \delta \leq \widehat{P}_t\left(s\right) \leq h_l\left(X_t\right) + \delta\right\}\right]\right)^{1/2}\right)^{1/2} \\
\leq \quad & \overline{ac}\left(\frac{1}{T}\sum_{t=1}^{T} \left(E_{\eta_t}\left[1\left\{h_l\left(X_t\right) - \delta \leq \widehat{P}_t\left(s\right) \leq h_l\left(X_t\right) + \delta\right\}\right]\right)^{1/2}\right)^{1/2} \\
= \quad & \overline{ac}\left(\frac{1}{T}\sum_{t=1}^{T} \left(E_{\eta_t}\left[E_{\eta_t}\left[1\left\{h_l\left(X_t\right) - \delta \leq \widehat{P}_t\left(s\right) \leq h_l\left(X_t\right) + \delta\right\} | X_t, Z_t\right]\right]\right)^{1/2}\right)^{1/2} \qquad (49)
\end{aligned}
$$

where $\triangle$ denotes the symmetric difference. The first inequality follows from Cauchy-Scwartz inequality; the second inequality, from Condition 7, taking $\max_{1\leq j\leq J_T} E\left[\left|p_j\left(Z_t\right)\right|^4\right] \leq \overline{c}$ and from

$F_j(Z_t) = \overline{a}\,|p_j(Z_t)|$, for some finite $\overline{a} \geq 1$. The last equality, from the law of iterated expectations. Note that

$$\begin{aligned} & 1\left\{h_l(X_t) - \delta \leq \widehat{P}_t(s) \leq h_l(X_t) + \delta\right\} \\ \leq\ & 1\left\{h_l(X_t) - \widehat{A}_T(s) - \delta \leq P_t(s) \leq h_l(X_t) + \widehat{A}_T(s) + \delta\right\} \end{aligned}$$

where $\widehat{A}_T(s) = \max_{t \leq T}\left|\widehat{P}_t(s) - P_t(s)\right|$. Therefore, by the mean-value theorem and Condition 8(i)

$$\begin{aligned} & E_{\eta_t}\left[1\left\{h_l(X_t) - \delta \leq \widehat{P}_t(s) \leq h_l(X_t) + \delta\right\}|X_t, Z_t\right] \\ \leq\ & E_{\eta_t}\left[1\left\{h_l(X_t) - \widehat{A}_T(s) - \delta \leq P_t(s) \leq h_l(X_t) + \widehat{A}_T(s) + \delta\right\}|X_t, Z_t\right] \\ =\ & F_{P_s|X,Z}\left(h_l(X_t) + \delta + \widehat{A}_T(s)\,|X_t, Z_t\right) \\ & -F_{P_s|X,Z}\left(h_l(X_t) - \delta - \widehat{A}_T(s)\,|X_t, Z_t\right) \\ =\ & \left\{\int_0^1 f_{P_s/X,Z}\left(h_l(X_t) + t\left(\delta + \widehat{A}_T(s)\right)|X_t, Z_t\right)dt\right\} \times \left[\delta + \widehat{A}_T(s)\right] \\ & +\left\{\int_0^1 f_{P_s/X,Z}\left(h_l(X_t) - t\left(\delta + \widehat{A}_T(s)\right)|X_t, Z_t\right)dt\right\} \times \left[\delta + \widehat{A}_T(s)\right] \\ \leq\ & 2K\left[\delta + \widehat{A}_T(s)\right]. \end{aligned}$$

So, the $\|\cdot\|_{2,T}$-size of the bracket is

$$\begin{aligned} & \|A_l \triangle B_l\|_{2,T} \\ \leq\ & \overline{ac}\left(\frac{1}{T}\sum_{t=1}^{T}\left(E_{\eta_t}\left[E_{\eta_t}\left[1\left\{h_l(X_t) - \delta \leq \widehat{P}_t(s) \leq h_l(X_t) + \delta\right\}|X_t, Z_t\right]\right]\right)^{1/2}\right)^{1/2} \\ \leq\ & \overline{ac}\left(\frac{1}{T}\sum_{t=1}^{T}\left(E_{\eta_t}\left[2K\delta + 2K\widehat{A}_T(s)\right]\right)^{1/2}\right)^{1/2} \\ \leq\ & \overline{ac}\left((2K\delta)^{1/2} + \left(2KE_{\eta^T}\left[\widehat{A}_T(s)\right]\right)^{1/2}\right)^{1/2} \\ \leq\ & \overline{ac}\,(2K)^{1/4}\left(\delta^{1/4} + \left(E_{\eta^T}\left[\widehat{A}_T(s)\right]\right)^{1/4}\right), \end{aligned}$$

where the last two inequalities apply the C_r-inequality.

Collect the constant $\overline{ac}\,(2K)^{1/4}$ into a single c. We conclude that

$$\log N_{[]}\left(c\left(\delta^{1/4} + \left(E_{\eta^T}\left[\widehat{A}_T(s)\right]\right)^{1/4}\right), \varrho_j, \|\cdot\|_{2,T}\right) \leq \log N\left(\delta, \mathcal{H}_K, \|\cdot\|_{\infty,\varpi}\right).$$

If we take $\omega = c\left(\delta^{1/4} + \left(E_{\eta^T}\left[\widehat{A}_T(s)\right]\right)^{1/4}\right)$, then

$$\log N_{[]}\left(\omega, \varrho_j, \|\cdot\|_{2,T}\right) \leq \log N\left(\left[\frac{\omega}{c} - \left(E_{\eta^T}\left[\widehat{A}_T(s)\right]\right)^{1/4}\right]^4, \mathcal{H}_K, \|\cdot\|_{\infty,\varpi}\right).$$

Therefore,

$$
\begin{aligned}
& \int_0^1 \sqrt{\log N_{[]}\left(\frac{E\left[F_j^2\right]}{4}\omega, \varrho_j, \|\cdot\|_{2,T}\right)} d\omega \\
\leq & \int_0^1 \sqrt{\log N\left(\left[\frac{E\left[F_j^2\right]}{4c}\omega - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{1/4}\right]^4, \mathcal{H}_K, \|\cdot\|_{\infty,\varpi}\right)} d\omega
\end{aligned}
$$

Next, use a change-in-variables argument. Let

$$
\nu = \left[\frac{E\left[F_j^2\right]}{4c}\omega - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{1/4}\right]^4
$$

Then[20]

$$
\begin{aligned}
if\ \omega &= 0 \Longrightarrow \nu = E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right] \equiv \nu_0 \\
if\ \omega &= 1 \Longrightarrow \nu = \left[\frac{E\left[F_j^2\right]}{4c} - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{1/4}\right]^4 \equiv \nu_1
\end{aligned}
$$

So,

$$
\begin{aligned}
& \int_0^1 \sqrt{\log N\left(\left[\frac{E\left[F_j^2\right]}{4c}\omega - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{1/4}\right]^4, \mathcal{H}_K, \|\cdot\|_{\infty,\varpi}\right)} d\omega \\
= & \int_{\nu_0}^{\nu_1} \sqrt{\log N\left(\nu, \mathcal{H}_K, \|\cdot\|_{\infty,\varpi}\right)} \left[\frac{c}{E\left[F_j^2\right]\nu^{3/4}}\right] d\nu \\
\leq & \left[\frac{c}{E\left[F_j^2\right]}\right] \int_{\nu_0}^{\nu_1} \sqrt{const.\nu^{-\frac{d_x}{\alpha}}} \left[\frac{1}{\nu^{3/4}}\right] d\nu \\
= & c \int_{\nu_0}^{\nu_1} \nu^{-\frac{(2d_x+3\alpha)}{4\alpha}} d\nu \\
= & c\left[\nu_1^{\frac{\alpha-2d_x}{4\alpha}} - \nu_0^{\frac{\alpha-2d_x}{4\alpha}}\right]
\end{aligned}
$$

[20] Note that to satisfy the inequality $\nu_1 \geq \nu_0$, we need the envelope function F_j to have a sufficiently large second moment. Formally, we need $\left[\frac{E\left[F_j^2\right]}{8\overline{ac}(2K)^{1/4}}\right]^4 \geq E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]$. Because $E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right] \leq 1$, it is sufficient to have $\frac{E\left[F_j^2\right]}{8\overline{ac}(2K)^{1/4}} \geq 1$. Remember that $F_j\left(Z_t\right) = \overline{a}\left|p_j\left(Z_t\right)\right|$, for some finite $\overline{a} \geq 1$, and that $\max_{1\leq j\leq J_T} E\left[\left|p_j\left(Z_t\right)\right|^4\right] \leq \overline{c}$. Then,

$$
\frac{E\left[F_j^2\right]}{8\overline{ac}\left(2K\right)^{1/4}} = \overline{a}\frac{E\left[p_j\left(Z_t\right)^2\right]}{8\overline{c}\left(2K\right)^{1/4}}.
$$

Even though $E\left[p_j\left(Z_t\right)^2\right] \leq \overline{c}$, we garantee the inequality $\nu_1 \geq \nu_0$ by taking a sufficiently large $\overline{a}$.

where c incorporates all constants. Next, use the definition of ν_0 and ν_1:

$$\begin{aligned}
& \left[\nu_1^{\frac{\alpha-2d_x}{4\alpha}} - \nu_0^{\frac{\alpha-2d_x}{4\alpha}}\right] \\
= \; & \left[\frac{E\left[F_j^2\right]}{4c} - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{1/4}\right]^{\frac{\alpha-2d_x}{\alpha}} - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{\frac{\alpha-2d_x}{4\alpha}} \\
\leq \; & \left[\left(\frac{E\left[F_j^2\right]}{4c}\right)^{\frac{\alpha-2d_x}{\alpha}} + \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{\frac{\alpha-2d_x}{4\alpha}}\right] - \left(E_{\eta^T}\left[\widehat{A}_T\left(s\right)\right]\right)^{\frac{\alpha-2d_x}{4\alpha}} \\
\leq \; & const.
\end{aligned}$$

where the first inequality uses the C_r-inequality: $\left(a+b\right)^r \leq c_r\left(a^r + b^r\right)$, where $c_r = 1$ if $r \leq 1$ and $c_r = 2^{r-1}$ if $r > 1$. We take $c_r = 1$ because $0 < \frac{(\alpha-2d_x)}{\alpha} < 1$, provided $\alpha > 2d_x$. So,

$$\int_0^1 \sqrt{\log N_{[]}\left(\frac{E\left[F_j^2\right]}{4}\omega, \varrho_j, \left\|\cdot\right\|_{2,T}\right)} d\omega \leq const.$$

■

A.2.2 Rate of Convergence

To obtain the rate of convergence we can concentrate our attention on a shrinking neighborhood around h_{0s}. Therefore, let $\mathcal{H}_{os} = \{h \in \mathcal{H} : \|h - h_0\|_{\infty,\omega} = o_p(1),\ \widehat{M}_T(h) \leq M_0\}$; and $\mathcal{H}_{osk} = \{h \in \mathcal{H}_k : \|h - \Pi_k h_0\|_{\infty,\omega} = o_p(1),\ \widehat{M}_T(h) \leq M_0\}$, where M_0 is a positive constant and $\Pi_k h_0$ is the projection of h_0 on $\mathcal{H}_k$. From now on, we work within these shrinking sets.

As usual, the rate of convergence is obtained under a weak norm first. We take the strong norm to be the $L^2\left(f_X\right)$ norm, $\left\|\cdot\right\|_{L_2(f_X)}$. The weak norm is obtained following the arguments in Chen and Pouzo (2012). Define the pathwise derivative at the direction $[h - h_0]$ evaluated at h_0:

$$\frac{dm(Z,h_0)}{dh}[h - h_0] \equiv \frac{dm(Z,(1-\tau)h_0 + \tau h)}{d\tau} \mid_{\tau=0} a.s.\mathcal{Z} \tag{50}$$

and define the weak norm as

$$\|h_1 - h_2\|^2 = E\left\{\left\|\frac{dm(Z,h_0)}{dh}[h_1 - h_2]\right\|_E^2\right\}. \tag{51}$$

So, in our case, from (20) and Condition 8, we have the well-defined norm

$$\|h_1 - h_2\|^2 = E\left\{E\left[f_{P_s|X,Z}\left(h_0(X,u)|X,Z\right)\left[h_1(X,u) - h_2(X,u)\right] \mid Z\right]\right\}^2.$$

To obtain the rate of convergence under the weak norm, we need it to be continuous with respect to the populational criterion function $E\left[m(Z,h)' m(Z,h)\right]$. The next condition is sufficient for that.

Condition 17 *Let c_0 and c_1 be finite positive constants. For all $h \in \mathcal{H}_{os}$ and almost all X, Z,*

$$0 < c_0 < \frac{f_{P_s|X,Z}\left(h(X,u)|X,Z\right)}{f_{P_s|X,Z}\left(h_0(X,u)|X,Z\right)} \leq c_1 < \infty.$$

For any $h \in \mathcal{H}_{os}$, define the linear integral operator

$$T_h\left[g\right] = E\left[f_{P_s|X,Z}\left(h(X,u)|X,Z\right)\left[g\right] \mid Z\right]$$

that maps a function $g \in L^2\left(\mathbb{R}^{d_X}, f_X\right)$ to a function in $L^2\left(\mathcal{Z}, f_Z\right)$. Note that $\left\|h_1 - h_2\right\|^2 = E\left\{T_{h_0}\left[h_1 - h_2\right]\right\}^2$. This operator is important to link the weak norm to the strong norm, so we can move from the rate of convergence under the weak norm to the rate under the strong norm. This requires that we work on Hilbert spaces.

Condition 18 *$\mathcal{H}_{K(T)}$ is a tensor product wavelet closed linear subspace of $\mathcal{H} \subset W_2^{\alpha}\left(\mathcal{X}, \omega\right)$. Let $\dim\left(\mathcal{H}_{K(T)}\right) = K_T < \infty$.*

The space $\mathcal{H}$ is a subset of the separable Hilbert space $\left(L^2\left(\mathbb{R}^{d_X}, f_X\right), \|\cdot\|_{L^2(f_X)}\right)$. $\mathcal{H}_{os}$ is contained in this Hilbert space and we take the sieves spaces $\mathcal{H}_K$ to be a (Riesz) basis of $L^2\left(\mathbb{R}^{d_x}, leb\right)$. We use

$$\mathcal{H}_{K(T)} = \left\{h \in W_2^{\alpha}\left(\mathcal{X}, \omega\right) : h\left(\cdot\right) = \sum_{k=1}^{K_T} a_k \psi_k\left(\cdot\right) \text{ and } 0 \leq h \leq 1\right\}$$

where $\left\{\psi_k : k \geq 1\right\}$ is a tensor product wavelet basis for $L^2\left(\mathbb{R}^{d_X}, f_X\right)$. Assume that

Condition 19 *(i) If $T_{h_{0s}}\left[g\right] = 0$, then $g = 0$ for all $g + h_{0s} \in \mathcal{H}_{os}$; and (ii) there is a non-negative, continuous increasing function φ such that*

$$\left\|T_{h_{0s}}\left[g\right]\right\|_{L_2(f_Z)}^2 \asymp \sum_{k=1}^{\infty} \varphi\left(k^{-2/d_x}\right) \left|\left\langle g, \psi_k\right\rangle_s\right|^2$$

for all $g \in \mathcal{H}_{os} \cap Domain\left(T_{h_{0s}}\right)$.

Condition 19(i) states that the new operator $T_{h_{0s}}$ is one-to-one and, so, preserves the identification of h_{0s}. Condition 19(ii) states that $\left\|T_{h_{0s}}\left[g\right]\right\|_{L_2(f_Z)}^2 = \|g\|^2$ is locally equivalent to $\sum_{k=1}^{\infty} \varphi\left(k^{-2/d_x}\right) \left|\left\langle g, \psi_k\right\rangle_s\right|^2$. Remember that $\|g\|_{L_2(f_X)}^2 = \sum_{k=1}^{\infty} \left|\left\langle g, \psi_k\right\rangle_s\right|^2$, by the definition of basis in Hilbert spaces, so the terms $\varphi\left(k^{-2/d_x}\right)$ link the strong and the weak norms. The functional form of φ depends upon whether we have a mildly or a severely ill-posed problem.

Proof of Proposition 5. From the proof of consistency, we obtained the following result:

$$\sup_{h\in\mathcal{H}_K}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{m}(Z_t,h)'\widehat{m}\left(Z_t,h\right)-m(Z_t,h)'m\left(Z_t,h\right)\right)\right]=O_p\left(\delta_{NT}^2\right) \tag{52}$$

where $\delta_{NT}^2=\max\left\{r_{NT},\frac{J_T}{T},b_{mT}^2\right\}$. Given the conditions imposed, this proposition is proved following the arguments in Chen and Pouzo's (2012) Corollary 5.1 and Proposition 6.2. The only difference here from their proof is that our term δ_{NT} includes the extra r_{NT} because of the first step estimator. But other than that, the argument is the same. ■

A.3 Probabilistic Framework

The probabilistic framework justifies (i) the assumption stating that the conditional distribution $\Pr\left(Y_{it}\leq y,S_{it}\leq s\mid \mathrm{t}\right)$ is known for each group t and that the equality $\Pr\left(Y_{it}\leq y,S_{it}\leq s\mid \mathrm{t}\right)=\Pr\left(Y_{it}\leq y,S_{it}\leq s\mid \sigma\left(C_t\right)\right)$ holds (Assumption 2); (ii) the condition stating that the dataset $\{Y_{it},S_{it}:i=1,...,N_t\}$ is i.i.d. conditional on common shocks C_t (Condition 1); (iii) the fundamental equation of this paper - equation (12) in the main text; and (iv) the condition imposing that the unfeasible sample $\{P_t(s),X_t,Z_t\}_{t=1}^{T}$ is i.i.d. across t (Condition 2).

Let γ_τ denote some unit in the population that belongs to the group τ. Denote the set of all groups by $\mathcal{T}$, and the set of all units in group τ by Γ_τ. Let $\mathcal{T}$ be a countable set and let Γ_τ be an arbitrary topological space. The set of all population units is given by $\Gamma=\otimes_{\tau\in\mathcal{T}}\Gamma_\tau$.

For a population unit $\gamma_\tau\in\Gamma_\tau$ in group $\tau\in\mathcal{T}$, the vector $S(\gamma_\tau,\tau)\in\mathcal{S}$ $(\subseteq\mathbb{R}^{d_S})$ denotes the observable individual-specific covariates and $\varepsilon(\gamma_\tau,\tau)\in\mathcal{E}$ $(\subseteq\mathbb{R}^{d_\varepsilon})$, the unobservable individual heterogeneity. Similarly, for a group $\tau\in\mathcal{T}$, the vector $X(\tau)\in\mathcal{X}$ $(\subseteq\mathbb{R}^{d_X})$ denotes the observable group-specific covariates and $U(\cdot,\tau)$, the unobservable "group-effect", which is a function mapping $\mathcal{S}$ into $\mathcal{U}$ $(\subseteq\mathbb{R})$. Assume $U\left(\cdot,\tau\right)\in\mathcal{J}\left(\mathcal{S}\right)$, where $\mathcal{J}\left(\mathcal{S}\right)$ denotes the set of functions mapping $\mathcal{S}$ to $\mathcal{U}$, and equip this space with a norm $\|\cdot\|_{\mathcal{J}}$ so that $\left(\mathcal{J}\left(\mathcal{S}\right),\|\cdot\|_{\mathcal{J}}\right)$ is a metric space.

Let $Y\left(\gamma_\tau,\tau\right)\in\mathcal{Y}$ $(\subseteq\mathbb{R})$ be the outcome of individual γ_τ in group τ, let $Y^*\left(\gamma_\tau,\tau\right)\in\mathcal{Y}^*$ $(\subseteq\mathbb{R})$ be the latent response variable and $Z\left(\tau\right)\in\mathcal{Z}$ $(\subseteq\mathbb{R}^{d_Z})$ be another group-level observable vector (the instrumental variables). Define the vectors

$$\begin{aligned}
V\left(\gamma_\tau,\tau\right) &= \left[Y\left(\gamma_\tau,\tau\right),Y^*\left(\gamma_\tau,\tau\right),S\left(\gamma_\tau,\tau\right),\varepsilon\left(\gamma_\tau,\tau\right)\right],\\
C\left(\tau\right) &= \left[X\left(\tau\right),U\left(\cdot,\tau\right),Z\left(\tau\right)\right],\text{ and}\\
W\left(\gamma_\tau,\tau\right) &= \left[V\left(\gamma_\tau,\tau\right),C\left(\tau\right)\right].
\end{aligned} \tag{53}$$

Let $(\Omega, \mathcal{F}, \mathcal{P})$ denote a probability space and $\omega \in \Omega$. For each unit $(\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}$, the vector $W(\gamma_\tau, \tau)$ is a random element defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. We assume the support of $W(\gamma_\tau, \tau)$ does not change with (γ_τ, τ) and is denoted by $\mathcal{W} \subseteq \mathcal{V} \times \mathcal{C}$, where $\mathcal{V} = \mathcal{Y} \times \mathcal{Y}^* \times \mathcal{S} \times \mathcal{E}$ and $\mathcal{C} = \mathcal{X} \times \mathcal{J}(\mathcal{S}) \times \mathcal{Z}$. We endow $\mathcal{W}$ with the (product) Borel sigma-field $\mathcal{A}$. Therefore, $W(\gamma_\tau, \tau, \omega)$ is an $\mathcal{F}$-measurable function mapping $(\Omega, \mathcal{F})$ into $(\mathcal{W}, \mathcal{A})$, i.e.,

$$W : \Gamma_\tau \times \mathcal{T} \times \Omega \to \mathcal{W}.$$

We endow $\mathcal{C}$ with the product Borel σ-field, denoted by $\mathcal{B}$, and, so, $C(\tau)$ maps $(\mathcal{W}, \mathcal{A})$ into $(\mathcal{C}, \mathcal{B})$. Denote by Q_τ the probability distribution (induced by $\mathcal{P}$) of the common shocks $C(\tau)$ on the space $(\mathcal{C}, \mathcal{B})$.

For each $(\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}$, the random $W(\gamma_\tau, \tau)$ satisfies the model restriction

$$\begin{aligned} Y(\gamma_\tau, \tau) &= D[Y^*(\gamma_\tau, \tau)] \\ Y^*(\gamma_\tau, \tau) &= G[S(\gamma_\tau, \tau), X(\tau), U(S(\gamma_\tau, \tau), \tau), \varepsilon(\gamma_\tau, \tau)] \end{aligned} \tag{54}$$

where $D(\cdot)$ is a known weakly increasing function and $G(\cdot)$ is a function defined on $\mathcal{S} \times \mathcal{X} \times \mathcal{U} \times \mathcal{E}$. It is assumed that the probability distribution of $\varepsilon(\gamma_\tau, \tau)$ (induced by $\mathcal{P}$) is independent of the other elements in $W(\gamma_\tau, \tau)$ and does not change with (γ_τ, τ).

Samples are obtained by drawing indices of units $\{\gamma_{i\tau} : i \geq 1\}$ randomly from Γ_τ according to some distribution $\mathcal{P}_{\Gamma_\tau}$ on Γ_τ. Because the groups themselves may be selected randomly we assume samples of groups are obtained by drawing indices $\{\tau_t : t \geq 1\}$ randomly from $\mathcal{T}$ according to some distribution $\mathcal{P}_\mathcal{T}$ on $\mathcal{T}$. The indices $\{\tau_t : t \geq 1\}$ and $\{\gamma_{i\tau} : i \geq 1\}$ are defined on $(\Omega, \mathcal{F}, \mathcal{P})$ the same probability space as $W(\gamma_\tau, \tau)$. More explicitly, we are assuming $\tau_t : \Omega \to \mathcal{T}$, and $\gamma_{i\tau} : \Omega \to \Gamma_\tau$. The units selected on Γ are therefore $\left\{\gamma_{i\tau_t} : i \geq 1, t \geq 1\right\}$, and, as a short cut notation, we let $\gamma_{it} = \gamma_{i\tau_t(\omega)}(\omega)$. The probability measure of γ_{it} is then $\mathcal{P}_\mathcal{T} \times \mathcal{P}_{\Gamma_\tau}$, the product of the probability of selecting a certain group and the probability of selecting a unit within the selected group.

Following Andrews (2005) we assume

Assumption A.1 *(i) $\{\gamma_{i\tau} : i \geq 1\}$ are i.i.d. indices independent of $\{W(\gamma_\tau, \tau) : (\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}\}$; and (ii) $\{\tau_t : t \geq 1\}$ are i.i.d. indices independent of $\{W(\gamma_\tau, \tau) : (\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}\}$.*

Assumption A.1 allows for proportional sampling by taking the distributions $\mathcal{P}_\mathcal{T}$ and $\mathcal{P}_{\Gamma_\tau}$ to be uniform on $\mathcal{T}$ and Γ_τ, respectively. But some units can be over-sampled when those distributions are not uniform. Still, the crucial restriction of Assumption A.1 is that it does not allow the sampling scheme to depend on the characteristics of the unit. In this sense, there is no sample selection here.

Given the framework and Assumption A.1, we justify the approach adopted in the main text in three steps: (i) first we state results valid for any group τ, justifying Assumption 2 (i.e., known $\Pr\left(Y_{it} \leq y, S_{it} \leq s \mid \sigma\left(C_t\right)\right)$ for each group t), Condition 1 (i.e., i.i.d. data conditional on $\sigma\left(C_t\right)$), and equation (12) in the main text. (ii) Then we introduce some randomness coming from randomly selecting groups $\tau_t\left(\omega\right)$, and show that the former results still hold. Finally, (iii) we obtain some properties of the distribution of $W\left(\gamma_{it}, \tau_t\right)$ across groups τ_t that justify Condition 2 (i.e., $\{P_t(s), X_t, Z_t\}_{t=1}^{T}$ i.i.d. across t). Items (i) and (ii) are important to justify the first step of our estimator - when we run regressions within groups - and item (iii) is important for the second step of our estimator - when we run regressions across groups.

A.3.1 Probabilistic Framework - Within Groups

For any group $\tau \in \mathcal{T}$, define $W_{i\tau} = W\left(\gamma_{i\tau}, \tau\right)$, and note that $\{W_{i\tau} : i \geq 1\}$ is a subordinated stochastic process (i.e., subordinated to $\{W\left(\gamma_\tau, \tau\right) : \gamma_\tau \in \Gamma_\tau\}$ via the directing process $\{\gamma_{i\tau} : i \geq 1\}$).[21] This process $\{W_{i\tau} : i \geq 1\}$ is defined on the probability space $\left(\mathcal{W}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \mathcal{P}_\tau^{\mathbb{N}}\right)$, where $\mathcal{W}^{\mathbb{N}}$ is the infinite product space, $\mathcal{A}^{\mathbb{N}}$ is the product Borel sigma-field on $\mathcal{W}^{\mathbb{N}}$ and $\mathcal{P}_\tau^{\mathbb{N}}$ is the probability measure on $\left(\mathcal{W}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}\right)$ induced by $\mathcal{P}$, by $\{W\left(\gamma_\tau, \tau\right) : \gamma_\tau \in \Gamma_\tau\}$ and by $\{\gamma_{i\tau} : i \geq 1\}$.

Given the sampling scheme specified in Assumption A.1(i), the random elements $\{W_{i\tau} : i \geq 1\}$ are exchangeable. That is, $\left(W_{\pi(1)\tau}, ..., W_{\pi(n)\tau}\right)$ has the same distribution as $\left(W_{1\tau}, ..., W_{n\tau}\right)$ for every permutation π of $(1, ..., n)$ for all $n \geq 2$. As a consequence, we can apply de Finetti's theorem [e.g., see Hall and Heyde (1980, Thm.7.1, p.203) and Andrews (2005)] and obtain the following Lemma:

Lemma 2 *Suppose Assumption A.1(i) holds. Then, for any $\tau \in \mathcal{T}$, $\{W_{i\tau} : i = 1, 2, ...\}$ are exchangeable and there exists a σ-field $\mathcal{G}_\tau \subset \mathcal{A}^{\mathbb{N}}$, such that, conditional on $\mathcal{G}_\tau$, $\{W_{i\tau} : i \geq 1\}$ are i.i.d..*

The σ-field $\mathcal{G}_\tau$ equals $\cap_{n=1}^{\infty}\mathcal{G}_{n,\tau}$, where $\mathcal{G}_{n,\tau}$ is the σ-field generated by n-symmetric functions that depend on $\{W_{i\tau} : i \geq 1\}$ and are invariant to permutations of the first n random elements of $\{W_{i\tau} : i = 1, ..., n\}$. This $\mathcal{G}_\tau$ is called the *symmetric* σ-field and equals the σ-field generated by the common elements of $\{W_{i\tau} : i \geq 1\}$. It is clear from the present context that the common elements are $C\left(\tau\right) = \left[X\left(\tau\right), U\left(\cdot, \tau\right), Z\left(\tau\right)\right]$. Hence, we have $\mathcal{G}_\tau = \sigma\left(C\left(\tau\right)\right)$, i.e., $\mathcal{G}_\tau$ is the σ-field generated

[21] See Feller (1966, Chap. X.7, p. 335).

by $C(\tau)$.[22]

Lemma 2 justifies Condition 1 in the main text (i.e., conditional i.i.d. observations within groups) if the groups are not randomly selected from $\mathcal{T}$. In cases where there is a sample of groups, we must go one step further and obtain a similar Lemma conditioned on the selected group. Before doing so, however, we note that an important consequence of this Lemma is that we can apply a Law of Large Numbers for exchangeable random variables, e.g., see Hall and Heyde (1980, (7.1), p. 202):

Lemma 3 *Suppose Assumption A.1(i) holds. Let* $r(\cdot)$ *be a vector-valued function with* $E\left\|r(W_{1\tau})\right\| < \infty$*, then, for any group* τ*,*

$$\frac{1}{N_\tau}\sum\nolimits_{i=1}^{N_\tau} r\left(W_{i\tau}\right) \rightarrow_p E\left[r(W_{1\tau})|\mathcal{G}_\tau\right] \qquad as\ N_\tau \rightarrow \infty,$$

where $\mathcal{G}_\tau$ *is the* σ*-field given in the Lemma 1.*

In particular, if we take $r\left(W_{i\tau}\right) = 1\left[Y_{i\tau} \leq y, S_{i\tau} \leq s\right]$, then it is possible to recover the joint distribution function of $(Y_{i\tau}, S_{i\tau})$ conditioned on the sub-sigma-field $\mathcal{G}_\tau = \sigma\left(C_\tau\right)$. Note however that it is not possible to recover the *unconditional* distribution of $(Y_{i\tau}, S_{i\tau})$, unless this vector is independent of $\sigma\left(C_\tau\right)$. This independence is ruled out by assumption and, as a result, all that can be known from the data within group τ is the conditional $\Pr\left(Y_{i\tau} \leq y, S_{i\tau} \leq s \mid \sigma\left(C_\tau\right)\right)$. That is what justifies Assumption 2(ii) (i.e., the conditional distribution $\Pr\left(Y_{it} \leq y, S_{it} \leq s \mid \mathrm{t}\right)$ is assumed to be known for each group $t = \tau$) and that the equality $\Pr\left(Y_{it} \leq y, S_{it} \leq s \mid \mathrm{t}\right) = \Pr\left(Y_{it} \leq y, S_{it} \leq s \mid \sigma\left(C_t\right)\right)$ holds (the equality (11) in Section 3).

Next, we justify the fundamental equation of the paper: equation (12). Because $Y_{i\tau}^*$ (and, so, $Y_{i\tau}$) is independent of Z_τ conditional on $(S_{i\tau}, X_\tau, U_\tau(\cdot))$, we have the equality

$$\Pr\left(Y_{i\tau} \leq y \mid S_{i\tau} = s, \sigma\left(C_\tau\right)\right) = \Pr\left(Y_{i\tau} \leq y \mid S_{i\tau} = s, \sigma\left(X_\tau, U_\tau(\cdot)\right)\right), \text{ a.s.},$$

where $\sigma\left(X_\tau, U_\tau(\cdot)\right) \subset \sigma\left(C_\tau\right)$ is the sub-sigma-field generated by $(X_\tau, U_\tau(\cdot))$. Moreover, we have

$$\Pr\left(Y_{i\tau} \leq y \mid S_{i\tau} = s, \sigma\left(X_\tau, U_\tau(\cdot)\right)\right) = \Pr\left(Y_{i\tau} \leq y \mid S_{i\tau} = s, \sigma\left(X_\tau, U_\tau\left(s\right)\right)\right), \text{ a.s.},$$

because conditioning on the sub-sigma-field $\sigma\left(X_\tau, U_\tau(\cdot)\right) \cap \{S_{i\tau} = s\}$ is equivalent to conditioning on the sub-sigma-field $\sigma\left(X_\tau, U_\tau(s)\right) \cap \{S_{i\tau} = s\}$.

[22] To be precise, the equality $\mathcal{G}_\tau = \sigma\left(C\left(\tau\right)\right)$ does not hold exactly. The correct statement is: (i) $\sigma\left(C\left(\tau\right)\right) \subset \mathcal{G}_\tau$, and (ii) if $A \in \mathcal{G}_\tau$, then there exists an element $B \in \sigma\left(C\left(\tau\right)\right)$ such that $A = B$, $\mathcal{P}_\tau^{\mathbb{N}}$-almost surely. For a proof see Meyer (1966, Lemma VIII-T2 and Theorem VIII-T3, p. 150) and note that his *symmetric* σ-field corresponds to our $\mathcal{G}_\tau$ and his *tail* σ-field $\mathcal{J}_{\infty,\tau} = \cap_{n\geq 1}\mathcal{J}_{n,\tau}$, where $\mathcal{J}_{n\tau} = \sigma\left(W_{n+1,\tau}, W_{n+2,\tau}, ...\right)$, corresponds to our $\sigma\left(C\left(\tau\right)\right)$. Then, Lemma 2 also holds with the sigma-field $\sigma\left(C\left(\tau\right)\right)$ in place of $\mathcal{G}_\tau$. Hence, without too much loss, we abuse notation and let $\mathcal{G}_\tau = \sigma\left(C\left(\tau\right)\right)$.

If we take $y = \widetilde{y}$, where $\widetilde{y}$ is the value referred in Assumption 1, and define $P_\tau(s, C_\tau) \equiv \Pr(Y_{i\tau} \leq \widetilde{y} \mid S_{i\tau} = s, \sigma(C_\tau))$, then

$$\begin{aligned} P_\tau(s, C_\tau) &\equiv \Pr(Y_{i\tau} \leq \widetilde{y} \mid S_{i\tau} = s, \sigma(C_\tau)) \\ &= \Pr(Y_{i\tau} \leq \widetilde{y} \mid S_{i\tau} = s, \sigma(X_\tau, U_\tau(s))), \text{ a.s..} \end{aligned} \tag{55}$$

This is the crucial equality (12) for the identification results. There are two important points to stress here. First, note that conditioning on $\{S_{i\tau} = s\}$ is essential to obtain the (almost sure) equality above, because $\sigma(X_\tau, U_\tau(s)) \subset \sigma(C_\tau)$ and, so, there is no guarantee that $\Pr(Y_{i\tau} \leq y \mid \sigma(C_\tau))$ would equal $\Pr(Y_{i\tau} \leq y \mid \sigma(X_\tau, U_\tau(s)))$ almost surely. Second, and related to the previous point, we are guaranteed to have i.i.d. data within group τ only when we condition on $\sigma(C_\tau)$, by Assumption A.1. Conditioning on the sub-sigma-field $\sigma(X_\tau, U_\tau(s))$ is not sufficient for i.i.d. data because the vector $(X_\tau, U_\tau(s))$ is *not* a common shock affecting all units in group τ.

Next, we introduce the randomness coming from the random selection of groups $\tau_t(\omega)$. To do so, define the random $W_{it} = W(\gamma_{it}, \tau_t)$. Let $\{W_{it} : i = 1, ..., N_t,\ t = 1, ..., T\}$ be the sample obtained by the method presented above. The model, in terms of the first sampled units, is

$$\begin{aligned} Y_{it} &= D(Y_{it}^*) \\ Y_{it}^* &= G(S_{it}, X_t, U_t(S_{it}), \varepsilon_{it}), \\ \text{for } i &= 1, ..., N_t,\ t = 1, ..., T \end{aligned} \tag{56}$$

and the sample data is given by $\{(Y_{it}, S_{it}, X_t, Z_t) : i = 1, ..., N_t,\ t = 1, ..., T\}$.

As before, we note that $\{W_{it} : i \geq 1\}$ is defined on $(\mathcal{W}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \mathcal{P}^{\mathbb{N}})$, but now $\mathcal{P}^{\mathbb{N}}$ is the probability measure on $(\mathcal{W}^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ induced by $\mathcal{P}$, by $\{W(\gamma_\tau, \tau) : (\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}\}$, by the indices $\{\gamma_{i\tau} : i \geq 1\}$ *and* by the indices $\{\tau_t : t \geq 1\}$.[23] Next, we establish the same result as in Lemma 2 but for the sequence $\{W_{it} : i \geq 1\}$.

Corollary 1 *Let $W_{it} = W_{i\tau_t}$ and let the σ-field $\mathcal{G}_t = \sigma(C(\tau_t)) \subset \mathcal{A}^{\mathbb{N}}$. Suppose Assumption A.1 holds. Then, for all $t \geq 1$, $\{W_{it} : i \geq 1\}$ are exchangeable and i.i.d. conditional on $\mathcal{G}_t$.*

Proof. *Conditional on the event $\{\tau_t(\omega) = \tau\} \in \mathcal{F}$, the probability measure of the sequence $\{W_{it} : i \geq 1\}$ is $\mathcal{P}_\tau^{\mathbb{N}}$. The indices $\{\tau_t : t \geq 1\}$ are drawn from the countable set $\mathcal{T}$ with distribution $\mathcal{P}_{\mathcal{T}}$. Therefore, the probability measure $\mathcal{P}^{\mathbb{N}}$ equals the product measure $\mathcal{P}_{\mathcal{T}} \times \mathcal{P}_\tau^{\mathbb{N}}$. Because $\mathcal{P}_\tau^{\mathbb{N}}$ is exchangeable for all $\tau \in \mathcal{T}$ from Lemma 2, the product $\mathcal{P}_{\mathcal{T}} \times \mathcal{P}_\tau^{\mathbb{N}}$ is exchangeable in the index i, and, so, the sequence $\{W_{it} : i \geq 1\}$ is exchangeable.*

[23] Now the process $\{W_{it} : i \geq 1\}$ is subordinated to $\{W(\gamma_\tau, \tau) : \gamma_\tau \in \Gamma_\tau\}$ using the directing process $\{\gamma_{i\tau_t}, \tau_t : i \geq 1, t \geq 1\}$.

Next, by applying Lemma 2 on the sequence $\{W_{it} : i \geq 1\}$ we conclude they are i.i.d. conditional on the sub-sigma-field generated by their common elements. In this case the common factors are $C_{\tau_t} = [X(\tau_t), U(\cdot, \tau_t), Z(\tau_t)]$, and we denote this σ-field by $\mathcal{G}_t = \sigma(C(\tau_t))$. ■

Let $\sigma(C(\tau_t)) = \sigma(C_t)$. Corollary 1 implies Condition 1 stated in Subsection 4.1 (i.e., conditional i.i.d. data for each group τ_t). We note that, because the groups are selected randomly from $\mathcal{T}$, the σ-fields $\sigma(C_\tau)$ and $\sigma(C_t)$ are not the same. Conditional on the event $\{\tau_t(\omega) = \tau\}$ we do have the equality $\sigma(C_\tau) = \sigma(C_t)$, but, unconditionally, $\sigma(C_\tau)$ is the sub-sigma-field generated by the random $C(\tau)$ with distribution Q_τ, while $\sigma(C_t)$ is the sub-sigma-field generated by the subordinated process $C(\tau_t)$ with distribution $Q = \mathcal{P}_{\mathcal{T}} \times Q_\tau$.

A.3.2 Probabilistic Framework - Across Groups

Next we consider the distribution of the random elements across groups τ_t. They are important to justify the second step of our estimator. First, we impose

Assumption A.2 *(i) For any $\tau \neq \tau' \in \mathcal{T}$, any $\gamma_\tau \in \Gamma_\tau$ and any $\widetilde{\gamma}_{\tau'} \in \Gamma_{\tau'}$, the random $W(\gamma_\tau, \tau)$ is independent of $W(\widetilde{\gamma}_{\tau'}, \tau')$; and (ii) the random $\{C(\tau) : \tau \in \mathcal{T}\}$ are identically distributed.*

Assumption A.2(i) states independence of the random $W(\gamma_\tau, \tau)$ across groups. There is therefore a fundamental asymmetry in the way we treat the observations within groups and across groups. We allow cross-sectional dependence within groups coming from the common shocks $C(\tau)$, but we do not allow dependence of $W(\gamma_\tau, \tau)$ across groups. It is possible to allow for dependence of $W(\gamma_\tau, \tau)$ across groups, but we do not pursue this approach here for simplicity. Moreover, we also restrict $C(\tau)$ to have identical distribution, so Q_τ is the same for all $\tau \in \mathcal{T}$.

Remember the definition $P_\tau(s, C_\tau) \equiv \Pr(Y_{i\tau} \leq \widetilde{y} \mid S_{i\tau} = s, \sigma(C_\tau))$, and define the analogous $P_t(s, C_t) \equiv \Pr(Y_{it} \leq \widetilde{y} \mid S_{it} = s, \sigma(C_t))$, for the cases where the groups are randomly selected. Assumption A.2 together with Assumption A.1 and Lemma 4 below imply that the unfeasible data $\{P_t(s), X_t, Z_t : t \geq 1\}$ is i.i.d.. If this data were feasible, we could use the results of Chen and Pouzo (2009, 2011) directly in the second step of our estimator. Despite this fact, Lemma 4 allows us to use some of Chen and Pouzo's results, which simplifies considerably the derivation of the asymptotic properties of our estimator.

Assumptions A.1 and A.2 imply the following Lemma:

Lemma 4 *Suppose Assumptions A.1 and A.2 hold. Then, $\{P_t(s, C_t), X_t, Z_t : t \geq 1\}$ are i.i.d., for any $s \in \mathcal{S}$.*

Proof. *Assumption A.2 states that* $\{C(\tau) : \tau \in \mathcal{T}\}$ *is i.i.d. across* τ*. So the sub-sigma-fields* $\{\mathcal{G}_\tau : \tau \in \mathcal{T}\}$ *are independent and* $\Pr(Y_{i\tau} \leq y, S_{i\tau} \leq s \mid \mathcal{G}_\tau)$ *also are independent across* τ*. Because* $\{S(\gamma_\tau, \tau) : (\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}\}$ *are independently distributed across* $\tau \in \mathcal{T}$ *by Assumption A.2(i), then, for any* $s \in \mathcal{S}$*,* $\Pr(Y(\gamma_\tau, \tau) \leq y \mid \{S(\gamma_\tau, \tau) = s\} \cap \mathcal{G}_\tau)$ *is independent of the conditional* $\Pr(Y(\gamma_{\tau'}, \tau') \leq y \mid \{S(\gamma_{\tau'}, \tau') = s\} \cap \mathcal{G}_{\tau'})$*, for any* $\tau \neq \tau'$*. In particular, if we take* y *to be* $\widetilde{y}$*, we conclude that* $P_\tau(s, C_\tau)$ *and* $P_{\tau'}(s, C_{\tau'})$ *are independent for any* $\tau \neq \tau'$*.*

Moreover, because $\{(\gamma_{it}, \tau_t) : i \geq 1, t \geq 1\}$ *are i.i.d. indices independent of {*$W(\gamma_\tau, \tau) : (\gamma_\tau, \tau) \in \Gamma_\tau \times \mathcal{T}$*} (by Assumption A.1), then, for any* $t \neq t'$*, the sub-sigma-fields* $\mathcal{G}_t \cap \{S(\gamma_{it}, \tau_t) = s\}$ *and* $\mathcal{G}_{t'} \cap \{S(\gamma_{it'}, \tau_{t'}) = s\}$ *must be independent of each other. We conclude* $P_t(s, C_t)$ *and* $P_{t'}(s, C_{t'})$ *also are independent for any* $t \neq t'$*.*

Next, because $P_\tau(s, C_\tau) = \Pr(Y_{i\tau} \leq \widetilde{y} \mid S_{i\tau} = s, X_\tau, U_\tau(s))$ *a.s., and because, for any* $\tau \neq \tau'$*,* $(X_\tau, U_\tau(s))$ *and* $(X_{\tau'}, U_{\tau'}(s))$ *are identically distributed by Assumption A.2, then, for all* $s \in \mathcal{S}$*,* $P_\tau(s, C_\tau)$ *and* $P_{\tau'}(s, C_{\tau'})$ *have the same distribution. The same result carries over to* $P_t(s, C_t)$ *and* $P_{t'}(s, C_{t'})$ *for any* $t \neq t'$ *by Assumptions A.1. Therefore,* $\{P_t(s, C_t), X_t, Z_t : t \geq 1\}$ *is i.i.d..* ■