A Time-Varying Endogenous Random Coefficient Model with an Application to Production Functions*

Ming Li†

September 23, 2021

Abstract

This paper proposes a random coefficient panel model where the regressors are correlated with the time-varying random coefficients in each period, a critical feature in many economic applications. We model the random coefficients as unknown functions of a fixed effect of arbitrary dimensions, a time-varying random shock that affects the choice of regressors, and an exogenous idiosyncratic shock. A sufficiency argument is used to control for the fixed effect, which enables one to construct a feasible control function for the random shock and subsequently identify the moments of the random coefficients. We propose a three-step series estimator and prove an asymptotic normality result. Simulation results show that the method can accurately estimate both the mean and the dispersion of the random coefficients. As an application, we estimate the average output elasticities for a sample of Chinese manufacturing firms.

Keywords: Unobserved heterogeneity, time-varying endogeneity, exchangeability, conditional control variable, production function estimation

*I thank Don Andrews, Yuichi Kitamura, Steve Berry, Joe Altonji, Xiaohong Chen, Yi Chen, JJ Forneron, Wayne Gao, Phil Haile, Kei Hirano, Mitsuru Igami, Sung Jae Jun, Xiaosheng Mu, Amil Petrin, Guangjun Shen, Tang Srisuma, Liangjun Su, Francis Vella, Ao Wang, Xinyang Wang, Weijie Zhong, and seminar participants at Georgetown, ITAM, NUS, PSU,THU, UGA and Yale for helpful comments. All errors are mine.

†Department of Economics, National University of Singapore, 1 Arts Link AS2, Singapore 117570, mli@nus.edu.sg.

1 Introduction

Linear panel models with fixed coefficients have been a workhorse in empirical research. A leading example concerns production function estimation, where the output elasticities with respect to each input are assumed to be the same both across firms and through time (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015). But it is neither theoretically proven nor empirically verified that the coefficients should be fixed. For example, why would Apple have the same capital elasticity as Sony? Moreover, why would Apple in 2019 have the same labor elasticity as in 2020 when almost everyone is working from home? Restricting the coefficients to be constant can lead to biased estimates of important model parameters such as output elasticity with respect to capital or labor (León-Ledesma, McAdam, and Willman, 2010), and consequently misguided policy recommendations, e.g., income distribution policy, tax policy, among others. Therefore, it is crucial to properly account for the unobserved heterogeneity both across individuals and through time in panel models.

To accommodate the rich forms of unobserved heterogeneity in the economy, one may consider linear panel models with random coefficients that are either independent of the regressors or satisfy certain distributional assumptions joint with or given the regressors (Mundlak, 1978; Chamberlain, 1984; Wooldridge, 2005a). However, because of the agent's optimization behavior, it is rarely the case that one can justify any ex-ante distributional assumptions on the joint distribution of the random coefficients and the regressors. To see this, consider a firm with individually unique and time-varying output elasticities with respect to each input. Then, in each period, the firm chooses inputs by maximizing its expected profits *after* taking the information about those heterogeneous elasticities into account. Consequently, such information enters its input choice decisions for each period in a potentially complicated way, making it very difficult, if not impossible, to put any distributional assumption on the joint distribution of the random coefficients and the regressors.

The combination of unobserved heterogeneity and correlation between the regressors and the time-varying random coefficients in each period poses significant challenges for the analyst. The fact that the information about the time-varying random coefficients is known to the agent when she optimally chooses the regressors but unobservable to the analyst gives rise to the classic simultaneity problem

(Marschak and Andrews, 1944). Allowing the regressors to depend on the unobserved (to the econometrician) time-varying random coefficients in each period in an unknown and potentially complicated way makes existing approaches not directly applicable (Chamberlain, 1992; Arellano and Bonhomme, 2012; Graham and Powell, 2012; Laage, 2020). Therefore, a new method is needed to deal with the challenges discussed so far to identify and estimate the parameters of interest, e.g., the average partial effects (APE) (Chamberlain, 1984; Wooldridge, 2005b).

This paper proposes a time-varying endogenous random coefficient (TERC) panel model where the regressors are correlated with the random coefficients in each period, a feature we call *time-varying endogeneity through the random coefficients*. The model is motivated by production function estimation, but can be applied to other important applications, e.g., labor supply estimation, Engel curve analysis, demand analysis, among many others (Blundell, MaCurdy, and Meghir, 2007b; Blundell, Chen, and Kristensen, 2007a; Chernozhukov, Hausman, and Newey, 2019). Specifically, the random coefficients in this paper are modeled as unknown and possibly nonlinear functions of a fixed effect of arbitrary dimensions, a random shock that captures per-period shocks to the agent, and an exogenous idiosyncratic shock which may or may be stationary. The modeling technique is based on the seminal paper of Graham and Powell (2012), with a major difference that will be discussed in detail in Section 2. Then, the regressors are determined by the agent's optimization behavior and expressed as unknown and possibly nonlinear functions of the fixed effect, random shock, and exogenous instruments. Consequently, the regressors are correlated with the random coefficients in each period.

For identification analysis, we first use an exchangeability assumption [1] on the conditional probability density function (pdf) of the vector of random shocks for all periods given the fixed effect to obtain a sufficient statistic for the fixed effect without parametric assumptions. Given the sufficient statistic, the agent's choice of regressors for any specific period is shown to not contain any additional information about the fixed effect. We exploit the independence result together with a monotonicity

[1] We acknowledge that the exchangeability condition limits how the random shocks can be correlated intertemporally, and discuss how to make it more flexible in Section 2. We also show how it is different from Altonji and Matzkin (2005)'s exchangeability condition, and why the rank condition can be more easily satisfied in our setting.

assumption to create a feasible conditional control variable [2] for the time-varying random shock. Finally, we adopt a sequential argument based on the independence result obtained in the first step, the feasible control variable constructed in the second step, and the law of iterated expectations (LIE), to identify the moments of the random coefficients. The intuition of the last step is that after conditioning on the sufficient statistic and the feasible conditional control variable, the residual variations in the regressors are exogenous. We further discuss several extensions to identify higher-order moments and allow for more flexible intertemporal correlations among the random shocks at the end of the section.

The constructive identification analysis leads to multi-step series estimators for both conditional and unconditional moments of the random coefficients. We derive convergence rates and prove asymptotic normality for the proposed estimators. The new inference results build on existing ones for multi-step series estimators (Andrews, 1991; Newey, 1997; Imbens and Newey, 2009; Lee, 2018; Hahn and Ridder, 2013, 2019). The main deviations from the literature include that the object of interest is a partial mean process (Newey, 1994) of the derivative of the second-step estimator with a nonseparable first step, and that the last step of the three-step estimation is an unknown but only estimable functional of the conditional expectation of the outcome variable. Thus, one needs to take the estimation error from each of the three steps into consideration to obtain correct large sample properties.

Simulation results show that the proposed method can accurately estimate both the mean and the dispersion of the random coefficients. The mean of the random coefficients has long been the central object of interest in empirical research as it measures on average how responsive the outcome is to exogenous changes in regressors. Depending on the conditioning variables, the distribution of the conditional moments of the random coefficients may also be useful to answer policy-related questions. For example, to what extent is a new labor-augmenting technology being diffused across firms? One may get an idea by looking at the dispersion of the estimated conditional moments of labor elasticities.

Finally, the procedure is applied to comprehensive panel data on the production process for Chinese manufacturing firms. Specifically, we estimate unconditional

[2] As will be explained in Section 3, to uniquely pin down the random shock one needs to fix the feasible control variable, the sufficient statistic, and the unobservable fixed effect simultaneously, which is different from classic control variables. Thus, we call the feasible control variable "conditional control variable."

expectations of the output elasticities, and compare them with the estimates using the method of Ackerberg, Caves, and Frazer (2015) on the same dataset. Then, we show the distribution of the conditional means of the elasticities for each sector and each year. Changes in both the mean and the dispersion of output elasticities are observed, which can be used to answer policy-relevant questions.

The rest of this paper is organized as follows. Section 2 introduces the model and key assumptions. Section 3 presents the identification strategy. Series estimators are provided in Section 4, together with their asymptotic properties. Section 5 contains a simulation study. In Section 6, we apply the method to panel data for the Chinese manufacturing firms to estimate their production functions. Finally, Section 7 concludes. All the proofs and an index of notation are presented in the Appendix.

1.1 Related Literature

We review the three lines of literature that this paper is connected to. The first line concerns random coefficient models. See Hsiao (2014) for a comprehensive survey. The closest paper to ours is Graham and Powell (2012), who consider the identification of the APE in a linear panel model with time-varying random coefficients. Compared with the celebrated paper by Chamberlain (1992) who considers regular identification and derives the semiparametric variance bound of the APE, Graham and Powell (2012) show that the APE is irregularly identified when the number of periods equals the dimension of the regressors via a novel DID type of argument. Later in Graham, Hahn, Poirier, and Powell (2018) they show how to identify quantiles of the random coefficients in a general linear quantile regression model. However, as explained in Section 2, the time stationarity assumption on the conditional distribution of random shocks given the whole vector of regressors in these two papers effectively rules out the time-varying endogeneity through the random coefficients. Therefore, their method does not directly apply here. Instead, we propose a different method for identification based on an exchangeability assumption and the control function approach.

Another closely related paper is Arellano and Bonhomme (2012), which considers a time-invariant random coefficient model. They exploit the information on the time dependence of the residuals to obtain identification of variances and distribution functions of the random coefficients. Their model assumptions and analysis are

very different from ours. Similarly to Arellano and Bonhomme (2012), Laage (2020) also considers a time-invariant random coefficient linear panel model with an additive scalar-valued fixed effect. As a result, her method does not apply.

The second line of research concerns the techniques used in this paper. The concept of exchangeable sequences dates back to Jonnson (1924), and has been used in many papers in economics (McCall, 1991; Kyriazidou, 1997; Altonji and Matzkin, 2005). The closest paper in this aspect to our work is Altonji and Matzkin (2005), who assume the conditional density of the fixed effect and random shock given the observable regressors for all periods is symmetric in the regressors. This assumption is not applicable to our setting, and we propose a different exchangeability condition on the unobservable variables.

This paper is also related to the literature on the control function approach in triangular systems (Newey, Powell, and Vella, 1999; Florens, Heckman, Meghir, and Vytlacil, 2008; Imbens and Newey, 2009; Kasy, 2011; Torgovitsky, 2015; D'Haultfœuille and Février, 2015). The construction of the feasible control variable for the random shock is built upon Imbens and Newey (2009), who assume a nonseparable first-step equation that determines the regressors and suggest a conditional cumulative distribution function (cdf) based approach for the identification. The main difference between our model and theirs is that in the first step we have two unobserved heterogeneity terms comprised of a fixed effect of arbitrary dimensions and a time-varying random shock, whereas Imbens and Newey (2009) assume one scalar-valued unobserved shock. Therefore, one cannot directly apply their method to our problem because the control variable constructed using their method will be infeasible. Instead, we use the implied conditional independence result from the sufficiency argument together with a monotonicity assumption to construct the conditional control variable for the random shock.

The third line of research concerns the production function estimation (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg, Caves, and Frazer, 2015). Differently from these influential works, we allow for the time-varying endogeneity through not only the random intercept (TFP), but also output elasticities modeled as time-varying random coefficients. Furthermore, we include a nonlinear fixed effect of arbitrary dimensions and propose a different identification strategy. Recently, several innovations (Kasahara, Schrimpf, and Suzuki, 2015; Li and Sasaki, 2017; Fox, Haddad, Hoderlein, Petrin, and Sherman, 2016; Gandhi, Navarro, and Rivers, 2020;

Demirer, 2020) have been made to relax the assumption of fixed output elasticities with respect to each input. The assumptions and method of this paper are very different from those mentioned above.

2 Model

In this section, we present the TERC model where the regressors can depend on the random coefficients in each period. We provide three applications that share this feature, followed by assumptions on model primitives.

Consider the following triangular simultaneous equations model with time-varying random coefficients:

$$Y_{it} = X_{it}^{'}\beta_{it} + \varepsilon_{it}, \tag{1}$$

$$\beta_{it} = \beta\left(A_i, U_{it}, U_{2,it}\right), \tag{2}$$

$$X_{it} = g\left(Z_{it}, A_i, U_{it}\right), \tag{3}$$

where:

- $i \in \{1, ..., n\}$ indexes n agents and $t \in \{1, ..., T\}$ indexes $T \geq 2$ periods.
- $Y_{it} \in \mathbb{R}$ represents the scalar-valued outcome variable for agent i in period t.
- $X_{it} \in \mathbb{R}^{d_X}$ is a vector of choice variables of the ith decision maker in period t.
- $Z_{it} \in \mathbb{R}^{d_Z}$ is a vector of exogenous instruments that affects the choice of X_{it} and is independent of $(A_i, U_{it}, U_{i,2t})$.
- A_i represents a fixed effect of arbitrary dimensions.
- $U_{it} \in \mathbb{R}$ is a scalar-valued shock that is continuously distributed and affects the choice of X_{it}. $U_{2,it}$ is a vector-valued idiosyncratic shock to β_{it} and is independent of all other variables.
- $\beta_{it} \in \mathbb{R}^{d_X}$ is a vector of random coefficients, the central object of interest. They are modeled as unknown and possibly nonlinear functions of $(A_i, U_{it}, U_{2,it})$.
- $\varepsilon_{it} \in \mathbb{R}$ is a scalar-valued error term independent of all other variables, with its mean normalized to zero. It can be considered as the measurement error or ex-post shocks.

- $g(\cdot)$ is a vector-valued function of (Z_{it}, A_i, U_{it}) that determines each coordinate of the choice variables X_{it}.

The modeling choice of β_{it} follows that of Graham and Powell (2012), with a difference that we will discuss later. We denote $\mathbf{X}_i = (X_{i1}, .., X_{iT})'$, $\mathbf{Z}_i = (Z_{it}, .., Z_{iT})'$, $Y_i = (Y_{i1}, .., Y_{iT})'$, $U_i = (U_{i1}, .., U_{iT})'$, $U_{2,i} = (U_{2,i1}, .., U_{2,iT})'$, $\varepsilon_i = (\varepsilon_{i1}, .., \varepsilon_{iT})'$.

To clarify the information structure of our model, first nature draws $(A_i, U_{it}, U_{2,it}, \varepsilon_{it})$ for each agent and period in the economy. Then, agent i in period t observes (A_i, U_{it}), which are correlated with β_{it} via (2). After knowing (A_i, U_{it}) and some exogenously determined instruments Z_{it}, agent i chooses X_{it} by solving some optimization problem (e.g., firm's profit maximization problem), leading to (3). Finally, agent i obtains its outcome Y_{it} via (1). We, as econometricians, observe $\{Y_{it}, X_{it}, Z_{it}\}$ for $i = 1, .., n$ and $t = 1, .., T$ as data and aim to recover distributional properties of β_{it}, such as $\mathbb{E}\beta_{it}$. We illustrate the information structure in Figure 1.

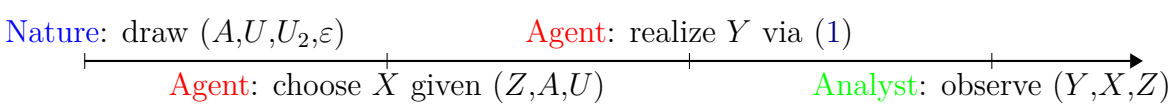

Figure 1: Information Structure

Model (1)–(3) arises in many economic applications. We mention a few here.

Example 1 (Production Function Estimation)**.** Suppose firm i in year t has production function in the classic C-D form of (1). As in (2), one can model β_{it} as a function of firm fixed effect A_i of arbitrary dimensions, a productivity shock U_{it} known to the firm, and an exogenous shock to the output elasticities $U_{2,it}$ (such as a macro shock that is relevant to the firm or measurement errors in its output elasticities) which may or may not be stationary over time. Suppose firm i observes (A_i, U_{it}) and some exogenously determined input prices Z_{it} such as interest rates and wages. Then, it chooses capital, labor and materials by solving a profit maximization problem using the information of (Z_{it}, A_i, U_{it}), leading to (3). Finally, firm i obtains its outcome Y_{it} via (1).

Example 2 (Labor Supply Estimation)**.** Suppose individual i has a linear labor supply function in the form of (1), where Y_{it} is the number of annual hours worked and X_{it} includes the endogenous hourly wage and other exogenous demographics. The

coordinate of β_{it} that corresponds to wage is the key object of interest as it quantifies how labor supply responds to wage rate variations over time. Then, given exogenous instruments Z_{it} such as county minimum wage or non-labor income, individual capability A_i, and random health shocks U_{it} to the individual, agent i chooses the job with a wage that solves her utility maximization problem, obtaining (3). Here, $U_{2,it}$ may capture measurement errors in her true ability and is thus exogenous to the choice of wage. See Blundell, MaCurdy, and Meghir (2007b) for more details on labor supply estimation.

Example 3 (Engel Curve Estimation)**.** Suppose the budget share of gasoline Y_{it} for household i at time t is a function of gas price and total expenditure in (1). Here $\beta_{it} \in \mathbb{R}^2$ is modeled as functions of the household fixed effect A_i, an idiosyncratic wealth shock U_{it} and exogenous shocks $U_{2,it}$ as in (2), and captures how elastic gasoline demand is with respect to total expenditure and gas price, respectively. Given (A_i, U_{it}) and an instrument of gross income of the head of household Z_{it}, household i optimally chooses its gas price and total expenditure budget by solving a utility maximization problem, obtaining (3). See Blundell, Chen, and Kristensen (2007a) for more details of the endogeneity issue in Engel curve estimation.

The time-varying correlation between X_{it} and β_{it} in these examples highlights the prevalence and importance of the time-varying endogeneity through the random coefficients. Nonetheless, to the best of our knowledge, existing literature has not yet incorporated this feature. Graham and Powell (2012) propose a panel model with time-varying random coefficients. Using their notation, they model $\beta_{it} = b^* (A_i, U_{it}) + d_t (U_{2,it})$ and assume $U_{2,it} \perp (\mathbf{X}_i, A_i)$. Note that we do not assume the additive structure in (2). The main difference between Graham and Powell (2012) and this paper is that they use a time stationarity assumption on the conditional distribution of U_{it} given $(\mathbf{X}_i, A_i)$:

$$U_{it} | \, \mathbf{X}_i, A_i \sim_d U_{is} | \, \mathbf{X}_i, A_i, \text{ for } t \neq s, \tag{4}$$

which effectively rules out time-varying endogeneity through the random coefficients. To see why, omit $U_{2,it}$ for now since it is exogenous. Consider a simple example where the number of periods $T = 2$ and the true data generating processes (dgp) of β_{it} and X_{it} are

$$\beta_{it} = A_i + U_{it}, \text{ and } X_{it} = \beta_{it}. \tag{5}$$

Suppose one observes $X_{i2} > X_{i1}$ in the data. It necessarily implies

$$\mathbb{E}\left[U_{i2} \middle| \mathbf{X}_i, A_i\right] > \mathbb{E}\left[U_{i1} \middle| \mathbf{X}_i, A_i\right], \tag{6}$$

thus violating (4). From this simple example, it is clear that under the time-stationarity condition (4) one cannot allow X_{it} to be correlated with β_{it} in each period such that one may infer distributional characteristics about β_{it} given $\mathbf{X}_i$. In contrast, we allow X_{it} to depend on U_{it} via (3). Similarly, Chernozhukov, Hausman, and Newey (2019) impose a time stationarity assumption on the conditional mean of the random coefficients given $\mathbf{X}_i$, again ruling out time-varying endogeneity through the random coefficients.

In addition to the issue of time-varying endogeneity through the random coefficients, our TERC model also features a nonseparable first step (3) that determines X_{it}, and a fixed effect A_i of arbitrary dimensions that enters the first step (3) possibly nonlinearly. The nonseparability of $g(\cdot)$ in the instrument Z_{it}, the fixed effect A_i, and the random shock U_{it} appears naturally due to the agent's optimization behavior[3]. There are two possible sources of nonlinearity in the fixed effect A_i: (i) the unknown random coefficients $\beta(A_i, U_{it}, U_{2,it})$ can be nonlinear in A_i, and (ii) the first-step equation $g(\cdot)$ due to agent optimization behavior can be nonlinear in A_i. Allowing a nonseparable first step $g(\cdot)$ and a nonlinear fixed effect A_i improves the flexibility of the model, however at the cost of greater analytical challenges for identification. For example, the usual demeaning or first differencing techniques no longer apply to the TERC model. We show how to deal with the fixed effect using a sufficiency argument.

It is worth mentioning that both A_i and U_{it} appear in the outcome equation (1) that determines Y_{it} and the first-step equation (3) that determines X_{it}. The modeling choice is motivated by economic applications since A_i and U_{it} play different roles in

[3]To see a simple example, omit ε_{it} in (1) and $U_{2,it}$ in (2) for now and model $\beta_{it} := \left(\beta_{it}^K, \beta_{it}^L, \omega_{it}\right)' = \beta(A_i, U_{it}) \in \mathbb{R}^3$. Then, given β_{it} and exogenous prices $Z_{it} = (p_{it}, r_{it}, w_{it})$, firm i solves

$$\max_{X_{it}=(K_{it},L_{it})} p_{it}Y_{it} - w_{it}L_{it} - r_{it}K_{it},$$

and obtains

$$\ln K_{it} = g^K(\beta_{it}, Z_{it}) := \left[\left(1-\beta_{it}^L\right)\ln\left(r_{it}/\beta_{it}^K\right) + \beta_{it}^L \ln\left(w_{it}/\beta_{it}^L\right) - \ln(\omega_{it}p_{it})\right] / \left(\beta_{it}^K + \beta_{it}^L - 1\right),$$
$$\ln L_{it} = g^L(\beta_{it}, Z_{it}) := \left[\left(1-\beta_{it}^K\right)\ln\left(w_{it}/\beta_{it}^L\right) + \beta_{it}^K \ln\left(r_{it}/\beta_{it}^K\right) - \ln(\omega_{it}p_{it})\right] / \left(\beta_{it}^K + \beta_{it}^L - 1\right).$$

Substituting $\beta_{it} = \beta(A_i, U_{it})$ into above equations gives (3).

β_{it} and agent i usually observes both, especially the individual fixed effect A_i. It is different from classic triangular simultaneous equations models (Newey, Powell, and Vella, 1999; Imbens and Newey, 2009) which assume in the equation that determines the choice of regressors there is only one unknown scalar that is arbitrarily correlated with (A_i, U_{it}). We consider (3) to be more realistic[4]. However, it makes identification challenging because now one has two unknown terms A_i and U_{it} in both (1) and (3). Thus, traditional control function approach does not directly apply. Instead, we show how to deal with (A_i, U_{it}) via a sequential argument in the next section.

Finally, it should be pointed out that the fixed effect A_i, modeled as an arbitrary dimensional object, effectively incorporates unobserved variations in the distributions of the idiosyncratic shocks U_{it}. For example, if the joint distribution of $(U_{i1}, .., U_{iT})$ is F_i which does not depend on time, then the whole function F_i can be incorporated as part of the fixed effect A_i, which may lie in a vector of infinite-dimensional functions. F_i captures a form of heteroskedasticity specific to each agent, and our method is robust to such forms of heterogeneity in error distributions without the need to specify F_i.

Next, we provide a list of assumptions on model primitives required for the subsequent identification analysis, and discuss them in relation to the model (1)–(3). We use U to denote U_{it} that affects the choice of X_{it} when there is no confusion, and write $U_{2,it}$ out completely.

Assumption 1 (Monotonicity of $g(\cdot)$). *At least one coordinate of $g(Z, A, U)$ is known to be strictly monotonic and continuously differentiable in U, for every realization of (Z, A) on its support.*

Assumption 1 is mild in the sense that it is satisfied in many applications and models. For example, in production function applications one may interpret U as the random R&D outcome. Then, the firm takes advantage of a positive shock (larger U) by purchasing more machines and hiring more workers, leading to a larger choice of capital and labor. Thus, Assumption 1 is satisfied. As in Newey, Powell, and Vella (1999), the assumption is automatically satisfied if $g(\cdot)$ is linear in U, but allows for

[4]One can argue that agent i may also know $U_{2,it}$ when choosing its X_{it}, which corresponds to the complete information setting in which agent i perfectly observes everything. We acknowledge this possibility. Due to the different interpretations of $(A_i, U_{it}, U_{2,it})$, we maintain the assumption that $U_{2,it}$ is independent of X_{it} in this paper.

more general forms of non-additive relations. An assumption similar to Assumption 1 is also imposed in Imbens and Newey (2009).

We note that strict monotonicity in U for all coordinates of g is not needed because a single scalar-valued U appears in both (1) and (3). If one has a model with a multi-dimensional U in (1) and each coordinate of U appearing in one equation of (3), then for the proposed method to work, all of the coordinates of g are required to be strictly monotonic in the corresponding coordinate of U. We discuss the vector-valued U case as an extension at the end of Section 3.

Assumption 2 (**Exchangeability**). *The conditional pdf of $U_{i1}, ..., U_{iT}$ given A_i wrt Lebesgue measure is continuous in $(u_{i1}, .., u_{iT})$ and exchangeable across t, i.e.*

$$f_{U_{i1},...,U_{iT}|A_i}\left(u_{i1}, ..., u_{iT}\middle| a_i\right) = f_{U_{i1},...,U_{iT}|A_i}\left(u_{it_1}, ..., u_{it_T}\middle| a_i\right), \tag{7}$$

where $(t_1, ..., t_T)$ is any permutation of $(1, ..., T)$.

Assumption 2 requires that the conditional density of U_i given A_i is invariant to any permutation of time. To see an example when it holds, suppose $T = 2$ and $U_{it} = A_i + \kappa_{it}$ for $t = 1, 2$, where κ_{it}'s are iid through time and independent of A_i. Then, Assumption 2 is satisfied and U_{i1} and U_{i2} are correlated via A_i. In this sense, Assumption 2 is milder than requiring U_{it} to be iid through time. Note that the simple example corresponds to the standard equicorrelated random effects specification due to Balestra and Nerlove (1966) from the panel analysis literature.

We emphasize that Assumption 2 requires exchangeability in the conditional density of U_{it}'s *given* A_i, thus allowing arbitrary correlation between A_i and U_{it}. For example, in production function estimation, one may expect that the better managerial capability a firm has, the greater chance a positive productivity shock shall occur. Such correlation is allowed under Assumption 2.

On the other hand, Assumption 2 does impose certain restrictions on how U_{it} can be correlated through time, which limits its applicability. For example, U_{it} cannot follow auto-regressive process under the assumption[5]. Nonetheless, given that A_i

[5]One way to allow for the random walk process among U_{it}'s is to replace (3) with

$$\Delta X_{it} = g\left(\Delta Z_{it}, A_i, \Delta U_{it}\right). \tag{8}$$

Then, one can rewrite Assumption 2 with ΔU_{it} in the position of U_{it}, thus allowing for the random walk process. The justification for (8), in the context of production function estimation, is that new

is allowed to be of arbitrary dimensions, one may include in A_i any distributional properties of U_i vector that are time-invariant (e.g., suppose the functional form of $\mathbb{E}[U_{it}|\, U_{it-1}]$ does not change over time then it can be included in A_i as well) and interpret U_{it} as a "normalized" idiosyncratic shock, which makes Assumption 2 more plausible to be satisfied. In addition, similarly as in Graham and Powell (2012), one can also extract trend related information from the exogenous shocks $U_{2,it}$. Finally, at the end of Section 3 we discuss several extensions that can allow for more flexible intertemporal correlations among U_{it}'s.

The use of exchangeability condition is not new. Altonji and Matzkin (2005) also impose an exchangeability assumption (Assumption 2.3 in their paper) to achieve identification in a nonparametric regression setting. Compared with their exchangeability condition, Assumption 2 avoids directly imposing distributional assumptions on the conditional density of U_{it} given the regressors $\mathbf{X}_i$ and is thus more primitive.

More precisely, Altonji and Matzkin (2005) denote $\Phi_{it} := (A_i, U_{it})$ and assumes

$$f_{\Phi_{it}|X_{i1},..,X_{iT}}\left(\varphi_{it}|\, x_{i1},..,x_{iT}\right) = f_{\Phi_{it}|X_{i1},..,X_{iT}}\left(\varphi_{it}|\, x_{it_1},..,x_{it_T}\right), \tag{9}$$

where $(t_1, ..., t_T)$ is any permutation of $(1, ..., T)$. There are two differences between (7) and (9). First, Altonji and Matzkin (2005) do not distinguish A_i from U_{it} in the definition of Φ_{it}, whereas A_i and U_{it} play different roles in this paper. The difference between A_i and U_{it} can be essential in applications such as production function estimation because they may have different economic interpretations and implications. Second, and more importantly, the exchangeability assumption (9) of Altonji and Matzkin (2005) requires the value of the conditional pdf of (A_i, U_{it}) given the regressors $(X_{i1}, .., X_{iT})$ does not depend on the order in which the regressors are entered into the function. In contrast, (7) requires that the conditional pdf of $(U_{i1}, .., U_{iT})$ given A_i is exchangeable in $(U_{it}, .., U_{iT})$, which is on the unknown primitives (A, U) rather than on (X, A, U) as in (9). Moreover, it could be challenging to justify (9) in our context since Φ_{it} includes U_{it} which determines X_{it} by (3), but not X_{is} for $s \neq t$, which creates asymmetry between X_{it} and X_{is} in (9). Note that the issue of the potential lack of variation in the regressors after conditioning on the sufficient

hiring or investment decision should be made based on changes in productivity because X_{it-1} already fully reflect the information contained in U_{it-1}. A sufficient condition for (8) is that $g(Z, A, U)$ is linear in Z and U.

statistic that appears in Altonji and Matzkin (2005) applies here too. We provide and discuss two conditions – one exploiting the linearity of (1), the other in the spirit of Altonji and Matzkin (2005) – in Assumption 6 to tackle the problem.

In light of the differences and observations discussed so far, we distinguish A_i from U_{it} in this paper and impose the exchangeability assumption on the conditional pdf of U_{it} given A_i in (7). In the next section, we use (7) to prove an exchangeability condition (11) on the conditional pdf of A_i given $(\mathbf{X}_i, \mathbf{Z}_i)$. We show that the derived exchangeability condition (11) guarantees the existence of a vector-valued function W_i symmetric in the elements of $(\mathbf{X}_i, \mathbf{Z}_i)$, such that conditioning on W_i the fixed effect A_i is independent of (X_{it}, Z_{it}) for any specific period t.

Assumption 3 (Random Sampling, Compact Support, and Exogeneity of $\boldsymbol{Z}$). *$(\mathbf{X}_i, \mathbf{Z}_i, Y_i, A_i, U_i, \varepsilon_i)$ is iid across $i \in \{1, ..., n\}$ with $n \to \infty$ and fixed $T \geq 2$. The support of (X_{it}, Z_{it}) is compact. $Z_{it} \perp (A_i, U_{it})$. ε_{it} and $U_{2,it}$ are independent of all other variables.*

The first part of Assumption 3 is a standard random sampling assumption[6]. The second part requires the support of (X_{it}, Z_{it}) to be compact, which we use to prove the sufficiency of W_i for A_i. The third part requires Z_{it} to be independent of (A_i, U_{it}). For example, in production function applications it is satisfied when the input market is perfectly competitive and Z_{it} are input prices. In the next section, we impose a conditional independence assumption between Z_{it} and (A_i, U_{it}) conditioning on W_i. Finally, we assume that ε_{it} and $U_{2,it}$ are random shocks independent of all other variables. Note that serial correlation is still allowed in ε_{it} and $U_{2,it}$ under Assumption 3.

3 Identification

In this section, we show how to identify the first-order moments of the random coefficient β_{it}. The analysis is divided into three steps. First, we obtain a sufficient statistic W_i for the fixed effect A_i via the exchangeability condition (7). Second, we construct a feasible variable V_{it} based on W_i and show that given (A_i, W_i), there is a

[6]One may include macro shocks common to all agents into the model and prove asymptotic results by using *conditional* law of large numbers and central limit theorems by conditioning on the σ-algebra generated by all of the random variables common to each individual i but specific to period t.

one-to-one mapping between V_{it} and U_{it}. Finally, we identify moments of β_{it} based on the LIE and the residual variation in X_{it} given (V_{it}, W_i).

Step 1: Sufficient Statistic for A_i

To construct a sufficient statistic for A_i, we exploit the exchangeability condition (7) and prove the following lemma.

Lemma 1 (**Sufficient Statistic for A_i**). *Suppose that Assumptions 1–3 are satisfied. Then, one can construct a feasible vector-valued function $W_i := W(\mathbf{X}_i, \mathbf{Z}_i)$ that is symmetric in the elements of $(\mathbf{X}_i, \mathbf{Z}_i)$ and satisfies*

$$f_{A_i|X_{it},Z_{it},W_i}(a_i|\, x_{it}, z_{it}, w_i) = f_{A_i|W_i}(a_i|\, w_i) \tag{10}$$

for any fixed $t \in \{1, ..., T\}$.

Lemma 1 exemplifies that one can exploit the panel data structure to control for complicated unobserved individual heterogeneity terms. The intuition of Lemma 1 is that W_i "absorbs" all the time-invariant information in the observable variables $\mathbf{X}_i$ and $\mathbf{Z}_i$. Given W_i, any t-specific X_{it} or Z_{it}, e.g., X_{i1}, Z_{i1}, does not contain any additional information about A_i. Therefore, one can exclude them from the conditioning set in (10) following the sufficiency argument. It is also worth emphasizing that Lemma 1 only concerns the pdf of the fixed effect A_i, not the random shock U_{it}.

To see an example of W_i, suppose $T = 2$ and both X_{it} and Z_{it} are scalars. Then, one example of such W_i is $T^{-1}\sum_t (X_{it}, Z_{it}, X_{it}^2, Z_{it}^2, X_{it}Z_{it})$. See Weyl (1939) for a detailed illustration on how to construct W_i. Notice that we do not impose any distributional assumption on the conditional density of A_i given (X_{it}, Z_{it}) in Lemma 1. With that said, ex-ante information about A_i can be incorporated to reduce the number of elements appearing in W_i. For example, when it is known that the probability distribution of A_i belongs to exponential family, such information can greatly simplify W_i. See Altonji and Matzkin (2005) for a more detailed discussion.

We prove Lemma 1 in Appendix A. The key to the proof involves a change of variables step that uses the exchangeability condition (7) to establish that the conditional density of A_i given $(\mathbf{X}_i, \mathbf{Z}_i)$ is exchangeable through time, i.e.,

$$f_{A_i|X_{i1},Z_{i1},..,X_{iT},Z_{iT}}(a_i|\, x_{i1}, z_{i1}, .., x_{iT}, z_{iT})$$

$$= f_{A_i|X_{i1},Z_{i1},..,X_{iT},Z_{iT}}\left(a_i|\, x_{it_1}, z_{it_1}, .., x_{it_T}, z_{it_T}\right), \tag{11}$$

where $(t_1, ..., t_T)$ is any permutation of $(1, ..., T)$. It is worth noting that the inclusion of Z_{it}'s in the conditioning set in (11) is necessary for the change of variable argument to go through. The exogeneity of Z_{it} is also crucial for the argument. Then, following Altonji and Matzkin (2005), one can construct a vector-valued function W_i symmetric in the elements of $(\mathbf{X}_i, \mathbf{Z}_i)$, using the Weierstrass approximation theorem and the fundamental theorem of symmetric functions, such that (10) hold.

Step 2: Feasible Control Variable for U_{it}

Given the nonseparable feature of the first-step $g(\cdot)$ function in (3), one may wish to use the method proposed in Imbens and Newey (2009) to construct a control variable for U_{it}. However, one cannot directly apply their technique in the current setting because we have two unknown heterogeneity terms A_i and U_{it}.

To see the issue more clearly, for brevity of exposition let X_{it} be a scalar that satisfies Assumption 1. Suppose one follows Imbens and Newey (2009) to construct a conditional CDF $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$, which under Assumption 1 equals $F_{U_{it}|A_i}(U_{it}|\, A_i)$, as the control variable for U_{it}. Then, two issues arise. First, $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$ is not feasible because A_i is unknown. Thus, it cannot be consistently estimated. Second, unlike the unconditional CDF $F_{U_{it}}(U_{it})$ derived in their setting which is a one-to-one mapping of U_{it}, the conditional CDF $F_{U_{it}|A_i}(U_{it}|\, A_i)$ is a function of both A_i and U_{it}. Therefore, one can not uniquely pin down U_{it} using $F_{U_{it}|A_i}(U_{it}|\, A_i)$ if A_i is unknown.

In this step, we deal with the first issue that $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$ is infeasible. The idea is to use the sufficient statistic W_i in Lemma 1 to get rid of A_i from the conditioning set of $F_{X_{it}|Z_{it},A_i}(X_{it}|\, Z_{it}, A_i)$. Specifically, the sufficiency condition (10) implies $A_i \perp (X_{it}, Z_{it})|\, W_i$, which further implies $X_{it} \perp A_i|\,(Z_{it}, W_i)$, i.e.,

$$f_{X_{it}|Z_{it},A_i,W_i}(x_{it}|\, z_{it}, a_i, w_i) = f_{X_{it}|Z_{it},W_i}(x_{it}|\, z_{it}, w_i). \tag{12}$$

The key observation here is the right hand side (rhs) of (12) is feasible since it only involves known or estimable objects from data. Suppose wlog the first coordinate of

X_{it} denoted by $X_{it}^{(1)}$ satisfies Assumption 1. Then, one can construct a feasible

$$V_{it} := F_{X_{it}^{(1)}\left|Z_{it},W_i\right.}\left(X_{it}^{(1)}\middle| Z_{it}, W_i\right) \tag{13}$$

and use (12) to deduce that

$$V_{it} = F_{X_{it}^{(1)}\left|Z_{it},W_i\right.}\left(X_{it}^{(1)}\middle| Z_{it}, W_i\right) = F_{X_{it}^{(1)}\left|Z_{it},A_i,W_i\right.}\left(X_{it}^{(1)}\middle| Z_{it}, A_i, W_i\right). \tag{14}$$

We add A_i back in (14) because we will use the monotonicity of $g(\cdot)$ in U_{it} to replace $X_{it}^{(1)}$ with U_{it} in $F_{X_{it}^{(1)}\left|Z_{it},A_i,W_i\right.}\left(X_{it}^{(1)}\middle| Z_{it}, A_i, W_i\right)$, which can only be done when A_i is also conditioned on.

Assumption 4 (**Conditional Independence**). $Z_{it} \perp U_{it}|\,(A_i, W_i)$.

Assumption 4 requires that the exogenous instrument Z_{it} is independent of U_{it} given (A_i, W_i). Since one may view W_i as summarizing all the time-invariant information about A_i in the data, the assumption is, loosely speaking, requiring $Z_{it} \perp U_{it}|\,A_i$, which is implied by $Z_{it} \perp (A_i, U_{it})$ in Assumption 3. When Assumption 4 is satisfied depends on the choice of W_i[7]. Assumption 4 is used to ensure that Z_{it} can be excluded from the conditioning set of $F_{U_{it}|Z_{it},A_i,W_i}(U_{it}|\,Z_{it}, A_i, W_i)$.

Lemma 2 (**Feasible Control Variable V_{it}**). *Suppose Assumptions 1–4 hold. Then, the random variable V_{it} satisfies*

$$V_{it} := F_{X_{it}^{(1)}\left|Z_{it},W_i\right.}\left(X_{it}^{(1)}\middle| Z_{it}, W_i\right) = F_{U_{it}|A_i,W_i}(U_{it}|\,A_i, W_i), \tag{15}$$

where $X_{it}^{(1)}$ denotes the first coordinate of X_{it} that is known to satisfy Assumption 1.

The important part of Lemma 2 is that V_{it} is feasible and can thus be consistently estimated. Note that having one coordinate of X_{it} that satisfies Assumption 1 is sufficient to construct V_{it}. When there are multiple coordinates of X_{it} that are known to satisfy Assumption 1, one can choose any coordinate of X_{it} to construct V_{it} because by (15), a single variable V_{it} suffices to control for the scalar-valued U_{it} given (A_i, W_i). We provide an extension when U_{it} is a vector towards the end of this section.

[7]For example, if X and Z are both scalars and one uses ex-ante information to choose $W_i = T^{-1}\sum_{t=1}^{T}\left(X_{it}, Z_{it}, X_{it}^2, Z_{it}^2\right)$ for $T = 2$, then a sufficient condition for Assumption 4 is that $g(Z_{it}, A_i, U_{it})$ is separable in Z_{it}.

However, the conditional CDF $F_{U_{it}|A_i,W_i}(U_{it}|\,A_i,W_i)$ on the rhs of (15) is not a one-to-one function of U_{it} because A_i is unknown. If A_i is known, then one can condition on (A_i, V_{it}, W_i), which by (15) is equivalent to conditioning on (A_i, U_{it}, W_i), and use the residual variation in X_{it} to identify moments of β_{it}. In the next step, we deal with unknown A_i using the sufficiency argument from the first step and the LIE.

Step 3: Identify the First-Order Moments of $\boldsymbol{\beta_{it}}$

We impose the next two regularity assumptions on $F_{U_{it}|A_i,W_i}(U_{it}|\,A_i,W_i)$ and the support of X_{it} given (V_{it}, W_i), respectively.

Assumption 5 (Strict Monotonicity of CDF of $\boldsymbol{U_{it}}$). *The conditional CDF $F_{U_{it}|A_i,W_i}(U_{it}|\,A_i,W_i)$ is strictly increasing in U_{it} for all (A_i, W_i).*

Assumption 5 requires that the conditional CDF of U_{it} given (A_i, W_i) cannot have flat areas, i.e., for each possible realization $c \in [0,1]$ of $F_{U_{it}|A_i,W_i}(U_{it}|\,A_i,W_i)$ and fixed (A_i, W_i), there is one and only one value of U_{it} such that $F_{U_{it}|A_i,W_i}(U_{it}|\,A_i,W_i) = c$. Consequently, fixing the level of $F_{U_{it}|A_i,W_i}(U_{it}|\,A_i,W_i)$ as well as (A_i, W_i) is equivalent to fixing the level of U_{it}.

Assumption 6 (Residual Variation in $\boldsymbol{X_{it}}$). *There are at least d_X linearly independent points in the support of X_{it} given V_{it} and W_i.*

Assumption 6 guarantees that there is sufficient variation in the support of X_{it} after conditioning on V_{it} and W_i. We need it to achieve identification of the moments of β_{it}, e.g., $\mathbb{E}\beta_{it}$. It is well known that it could be challenging to obtain enough residual variation in X_{it} given the sufficient statistic W_i. However, we are able to leverage the linearity in the outcome equation (1), which makes Assumption 6 easier to be satisfied. First, W_i only involves functions of the coordinates of X_{it} that are endogenous. When there are exogenous variables also contained in X_{it}, we can use the variation in the exogenous variables to achieve identification[8]. Second, one can use the symmetry in the solution to $W_i = \overline{w}$ and obtain at least T points that X_{it} vector can take. Thus, the longer panel we have, the larger the set of points that

[8]To see a concrete example, suppose $X_{it} \in \mathbb{R}^2$ and $X_{it}^{(1)}$ is endogenous while $X_{it}^{(2)}$ is exogenous. Suppose the conditional support of $X_{it}^{(1)}$ given (V_{it}, W_i) only contains a singleton. Then, as long as the *unconditional* support of $X_{it}^{(2)}$ contains more than 2 different points, one can follow the argument in this paper and identify moments of β_{it}.

satisfy $W_i = \overline{w}$ we may obtain[9]. Combining these two facts, we consider Assumption 6 to be a mild condition in our setup.

Next, we provide another assumption on the residual variation in X_{it} given (V_{it}, W_i) similarly to assumption 2.2 of Altonji and Matzkin (2005). It is stronger than Assumption 6, and usually requires ex-ante information on W_i for it to be satisfied. The flip side of it is that it enables us to use a derivative-based argument to identify and estimate moments of β_{it}, which we follow in the subsequent analysis.

Assumption 6′ (More Variation in X_{it}). *The support of X_{it} given V_{it} and W_i contains some ball of positive radius a.s. wrt (V_{it}, W_i).*

Suppose A_i is known for now, we have

$$\begin{aligned}
&\mathbb{E}\left[\beta_{it} \middle| X_{it}, A_i, V_{it}, W_i\right] \\
&= \mathbb{E}\left[\beta\left(A_i, U_{it}, U_{2,it}\right) \middle| g\left(Z_{it}, A_i, U_{it}\right), A_i, F_{U_{it}|A_i,W_i}\left(U_{it} \middle| A_i, W_i\right), W_i\right] \\
&= \mathbb{E}\left[\beta\left(A_i, U_{it}, U_{2,it}\right) \middle| A_i, V_{it}, W_i\right] =: \widetilde{\beta}\left(A_i, V_{it}, W_i\right), \qquad (16)
\end{aligned}$$

where the first equality holds by the definition of V_{it} and (3), and the second equality is true because the σ-algebra generated by $\left(A_i, F_{U_{it}|A_i,W_i}\left(U_{it} \middle| A_i, W_i\right), W_i\right)$ is equal to that generated by (A_i, U_{it}, W_i) by Assumption 5, which contains all the information necessary to calculate the first-order moment of β_{it} as a function of $(A_i, U_{it}, U_{2,it})$.

Next, to deal with unknown A_i appearing in (16), we use the LIE together with the sufficiency condition of (10). Specifically, taking the conditional expectation of $\widetilde{\beta}\left(A_i, V_{it}, W_i\right)$ wrt A_i given (X_{it}, V_{it}, W_i) gives

$$\begin{aligned}
&\mathbb{E}\left[\widetilde{\beta}\left(A_i, V_{it}, W_i\right) \middle| X_{it}, V_{it}, W_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\widetilde{\beta}\left(A_i, V_{it}, W_i\right) \middle| X_{it}, Z_{it}, W_i\right] \middle| X_{it}, V_{it}, W_i\right] \\
&= \mathbb{E}\left[\int \widetilde{\beta}\left(a, V_{it}, W_i\right) f_{A_i|W_i}\left(a \middle| W_i\right) \mu\left(da\right) \middle| X_{it}, V_{it}, W_i\right] =: \beta\left(V_{it}, W_i\right), \qquad (17)
\end{aligned}$$

[9]For example, when $T = 2$ and X_{it} is a scalar, we can construct $W_i = \left(X_{i1} + X_{i2}, X_{i1}^2 + X_{i2}^2\right)$ and show it satisfies the assumptions for Lemma 1. Then, by the symmetry of W_i, if $(X_{i1} = a, X_{i2} = b)$ is the solution to $W_i = \overline{w}$, then $(X_{i1} = b, X_{i2} = a)$ must also be a solution. Therefore, given $W_i = \overline{w}$ there are two possible values that X_{i1} can take. Note that depending on the specific DGP there is also a possibility that $a = b$, but with a larger T the likelihood that all X_{it} are the same becomes smaller.

where the first equality holds by the LIE and the fact that V_{it} is a function of (X_{it}, Z_{it}, W_i) and the second equality holds by (10). The measure $\mu(\cdot)$ in the third line of (17) represents the Lebesgue measure.

Given (17), taking the conditional expectation of both sides of (1) given (X_{it}, V_{it}, W_i) leads to

$$\mathbb{E}\left[Y_{it} \middle| X_{it}, V_{it}, W_i\right] = X_{it}^{'}\beta\left(V_{it}, W_i\right). \tag{18}$$

The result is intuitive because V_{it} is a feasible control variable for U_{it} given (A_i, W_i) and W_i is a sufficient statistic for A_i. Therefore, fixing (V_{it}, W_i) effectively controls for (A_i, U_{it}), and the residual variation in X_{it} becomes exogenous.

When Assumption 6′ holds[10], one can identify $\beta\left(V_{it}, W_i\right)$ by

$$\beta\left(V_{it}, W_i\right) = \partial\mathbb{E}\left[Y_{it} \middle| X_{it}, V_{it}, W_i\right]/\partial X_{it}. \tag{19}$$

Then, one can identify $\mathbb{E}\left[\beta_{it} \middle| X_{it}\right]$ and $\mathbb{E}\beta_{it}$ via the LIE. For example,

$$\mathbb{E}\beta_{it} = \mathbb{E}\beta\left(V_{it}, W_i\right) = \mathbb{E}\left(\partial\mathbb{E}\left[Y_{it} \middle| X_{it}, V_{it}, W_i\right]/\partial X_{it}\right), \tag{20}$$

where the expectation is taken wrt the joint distribution of (V_{it}, W_i), an identifiable object from data.

Theorem 1 (**Identification**). *If Assumptions 1–5 and either Assumption 6 or 6′ are satisfied, then* $\mathbb{E}\left[\beta_{it} \middle| V_{it}, W_i\right]$, $\mathbb{E}\left[\beta_{it} \middle| X_{it}\right]$, *and* $\mathbb{E}\beta_{it}$ *are identified.*

Theorem 1 presents the main identification result following the steps above. The idea is simple: find the feasible variables (V_{it}, W_i) such that conditioning on these variables, the residual variation in X_{it} is exogenous to that in β_{it}. The sufficient statistic W_i for A_i constructed in the first step plays an important role. It not only enables the construction of the feasible control variable V_{it} for U_{it} given (A_i, W_i) in the second step, but also manages to control for A_i in the last step. By exploiting the panel data structure, the proposed method extends the classic control function approach (Newey, Powell, and Vella, 1999; Imbens and Newey, 2009) to the setting with a fixed effect of arbitrary dimensions and a random shock, both of which affect the choice of X_{it} in a nonseparable way.

[10] When Assumption 6 is satisfied, one can identify $\beta\left(V_{it}, W_i\right)$ by solving a system of linear equations.

Higher-Order Moments of β_{it}

We have shown the identification of the first-order expectation of β_{it}. Higher-order moments such as variance of the random coefficients can also be of interest to researchers and policy makers. For example, policy makers may be interested in how fast labor-augmenting technology is being diffused among firms. In this section, we briefly discuss how to identify the second-order moments under regularity conditions.

For simplicity of exposition, we consider the case when the vector of regressors $(X_{it}, 1)$ is two-dimensional. With a slight abuse of notation, let $(\beta_{it}, \omega_{it}) \in \mathbb{R}^2$ where β_{it} is the random coefficient corresponding to the scalar X_{it} and ω_{it} is the random coefficient associated with the constant 1. The ex-post shock ε_{it} is omitted from the analysis for brevity[11].

Since (17) holds with β_{it}^2 or ω_{it}^2 in place of β_{it}, one has

$$\begin{aligned}
&\mathbb{E}\left[\beta_{it}^2 \middle| X_{it}, V_{it}, W_i\right] = \mathbb{E}\left[\beta_{it}^2 \middle| V_{it}, W_i\right], \\
&\mathbb{E}\left[\omega_{it}^2 \middle| X_{it}, V_{it}, W_i\right] = \mathbb{E}\left[\omega_{it}^2 \middle| V_{it}, W_i\right], \\
&\mathbb{E}\left[\omega_{it}\beta_{it} \middle| X_{it}, V_{it}, W_i\right] = \mathbb{E}\left[\omega_{it}\beta_{it} \middle| V_{it}, W_i\right].
\end{aligned} \tag{21}$$

Thus, taking the conditional expectation of the squares of both sides of (1) given (X_{it}, V_{it}, W_i) gives

$$\mathbb{E}\left[Y_{it}^2 \middle| X_{it}, V_{it}, W_i\right] = X_{it}^2\mathbb{E}\left[\beta_{it}^2 \middle| V_{it}, W_i\right] + 2X_{it}\mathbb{E}\left[\beta_{it}\omega_{it} \middle| V_{it}, W_i\right] + \mathbb{E}\left[\omega_{it}^2 \middle| V_{it}, W_i\right]. \tag{22}$$

Then, one can identify $\mathbb{E}\left[\beta_{it}^2 \middle| V_{it}, W_i\right]$ by exploiting the second-order derivative of $\mathbb{E}\left[Y_{it}^2 \middle| X_{it}, V_{it}, W_i\right]$ wrt X_{it}

$$\mathbb{E}\left[\beta_{it}^2 \middle| V_{it}, W_i\right] = \left(\partial^2\mathbb{E}\left[Y_{it}^2 \middle| X_{it}, V_{it}, W_i\right] / \partial X_{it}^2\right)/2, \tag{23}$$

and identify $\mathbb{E}\left[\beta_{it}\omega_{it} \middle| V_{it}, W_i\right]$ by

$$\mathbb{E}\left[\beta_{it}\omega_{it} \middle| V_{it}, W_i\right] = \left(\partial\mathbb{E}\left[Y_{it}^2 \middle| X_{it}, V_{it}, W_i\right] / \partial X_{it} - 2X_{it}\mathbb{E}\left[\beta_{it}^2 \middle| V_{it}, W_i\right]\right)/2. \tag{24}$$

[11] If ε_{it} is present, one may follow Arellano and Bonhomme (2012) to impose an ARMA structure on the intertemporal dependence among ε_{it}'s to identify the second-order moments of β_{it} and ω_{it}.

Finally, we identify $\mathbb{E}\left[\omega_{it}^2 \middle| V_{it}, W_i\right]$ by (22)

$$\begin{aligned}
&\mathbb{E}\left[\omega_{it}^2 \middle| V_{it}, W_i\right] \\
&= \mathbb{E}\left[Y_{it}^2 \middle| X_{it}, V_{it}, W_i\right] - X_{it}^2 \mathbb{E}\left[\beta_{it}^2 \middle| V_{it}, W_i\right] - 2X_{it}\mathbb{E}\left[\beta_{it}\omega_{it} \middle| V_{it}, W_i\right]. \qquad (25)
\end{aligned}$$

By induction, the analysis can be extended to identify any order of moments of β_{it}, which under regularity conditions (Stoyanov, 2000) uniquely determines the distribution function of β_{it}.

The flexible identification argument can also be used to identify intertemporal correlations of the random coefficients. For example, one can identify $\mathbb{E}\left[\beta_{it}\beta_{is} \middle| X_{it}, X_{is}, V_{it}, V_{is}, W_i\right]$ from $\mathbb{E}\left[Y_{it}Y_{is} \middle| X_{it}, X_{is}, V_{it}, V_{is}, W_i\right]$ for any $t, s \in \{1, .., T\}$ following an almost identical argument as in (21)–(25).

Other Extensions

The identification argument is flexible and can adapt to several extensions. First, when there is a vector of U_{it} (say two dimensional) in (1) while each coordinate of U_{it} appears in only one of (3), i.e.,

$$\begin{aligned}
Y_{it} &= X_{it}'\beta\left(A_i, U_{it}^{(1)}, U_{it}^{(2)}\right) + \varepsilon_{it} \\
X_{it}^{(1)} &= g^{(1)}\left(Z_{it}, A_i, U_{it}^{(1)}\right) \\
X_{it}^{(2)} &= g^{(2)}\left(Z_{it}, A_i, U_{it}^{(2)}\right),
\end{aligned}$$

one can construct

$$V_{it}^{(1)} := F_{X_{it}^{(1)}\middle| Z_{it}, W_i}\left(X_{it}^{(1)} \middle| Z_{it}, W_i\right) \text{ and } V_{it}^{(2)} := F_{X_{it}^{(2)}\middle| Z_{it}, W_i}\left(X_{it}^{(2)} \middle| Z_{it}, W_i\right),$$

and follow Step 1–3 to obtain

$$\mathbb{E}\left[Y_{it} \middle| X_{it}, V_{it}^{(1)}, V_{it}^{(2)}, W_i\right] = X_{it}'\beta\left(V_{it}^{(1)}, V_{it}^{(2)}, W_i\right).$$

Then, the identification follows identically to (19).

Second, to allow more flexible or even arbitrary intertemporal correlation than (7) among the U_{it}'s, one may replace the individual fixed effect A_i with a group fixed

effect A_j when i belongs to group j (Cameron, Gelbach, and Miller, 2012; Cameron and Miller, 2015). More precisely, we modify the model (1)–(3) to be

$$\begin{aligned} Y_{ijt} &= X_{ijt}^{'}\beta\left(A_j, U_{ijt}\right) + \varepsilon_{ijt}, \\ X_{ijt} &= g\left(Z_{ijt}, A_j, U_{ijt}\right), \end{aligned} \tag{26}$$

where j is the group that agent i belongs to and t represents time. One may want to use this model instead of (1)–(3) if she desires to relax the restriction on the intertemporal correlations between U_{it}'s and finds the evidence of a group fixed effect, e.g., location or sector or age fixed effect. Suppose that group j contains individuals $\{i_1, .., i_I\}$ and $I \leq n$. Let $U_{ij} = \left(U_{ij1}, ..., U_{ijT}\right)^{'}$. Then, one can use a "group" version of the exchangeability condition

$$f_{U_{i_1 j},...,U_{i_I j}|A_j}\left(u_{i_1 j}, ..., u_{i_I j}\middle| a_j\right) = f_{U_{i_1 j},...,U_{i_I j}|A_j}\left(u_{i_1^{'} j}, ..., u_{i_I^{'} j}\middle| a_j\right), \tag{27}$$

where $\left(i_1^{'}, ..., i_I^{'}\right)$ is any permutation of $(i_1, ..., i_I)$, to construct a sufficient statistic W_j for A_j and proceed as in Step 2–3 to identify moments of the random coefficients.

Third, to deal with persistent shocks to X_{it} or deterministic time trend in X_{it}, one may model the intertemporal change in X_{it}, or $\Delta X_{it} := X_{it} - X_{it-1}$, as a function g of (Z, A, U) in (3). The analysis is mostly the same as before, except that W_i is now a symmetric function in the elements of ΔX_{it} rather than X_{it} and $V_{it} := F_{\Delta X_{it}|Z_{it},W_i}$. Then, one can identify the moment of β_{it} by taking partial derivative wrt ΔX_{it} on both sides of

$$\mathbb{E}\left[Y_{it}\middle| X_{it-1}, \Delta X_{it}, V_{it}, W_i\right] = \left(X_{it-1} + \Delta X_{it}\right)^{'}\beta\left(X_{it-1}, V_{it}, W_i\right). \tag{28}$$

4 Estimation and Large Sample Theory

The constructive identification argument leads to a feasible estimator for the first-order moment of β_{it}. In this section, we first estimate the conditional and unconditional moments of the random coefficients using multi-step series estimators. Then, we obtain the convergence rates and asymptotic normality results for the proposed estimators.

4.1 Estimation

The parameters of interest considered here are

$$\beta(v,w) := \mathbb{E}\left[\beta_{it} \middle| V_{it} = v, W_i = w\right], \ \beta(x) := \mathbb{E}\left[\beta_{it} \middle| X_{it} = x\right], \text{ and } \overline{\beta} := \mathbb{E}\beta_{it}. \quad (29)$$

We propose to estimate them using three-step series estimators. In the first step, we estimate $V(x,z,w) = F_{X_{it}|Z_{it},W_i}(x|\, z,w)$ and denote $V_{it} := V(X_{it}, Z_{it}, W_i)$. Then, for $s = (x,v,w)$ we estimate $G(s) := \mathbb{E}\left[Y_{it} \middle| X_{it} = x, V_{it} = v, W_i = w\right]$ using $\widehat{V}$ obtained in the first step and denote $G_{it} := G(S_{it}) = G(X_{it}, V_{it}, W_i)$. Finally, we estimate $\beta(v,w)$, $\beta(X_{it})$ and $\overline{\beta}$, all of which are identifiable functionals of $G(s)$. For brevity of exposition, we provide definitions of all of the symbols appearing in this section in Appendix C.

More specifically, we first estimate $V(x,z,w)$ by regressing $\mathbb{1}\{X_{it} \leq x\}$ on the basis functions $q^L(\cdot)$ of (Z_{it}, W_i) with trimming function $\tau(\cdot)$:

$$\begin{aligned}
\widehat{V}(x,z,w) &= \tau\left(\widehat{F}_{X_{it}|Z_{it},W_i}(x|\, z,w)\right) \\
&= \tau\left(q^L(z,w)' \widehat{Q}^{-1} \sum_{j=1}^{n} q_j \mathbb{1}\{X_{jt} \leq x\}/n\right) \\
&=: \tau\left(q^L(z,w)' \widehat{\gamma}^L(x)\right). \qquad (30)
\end{aligned}$$

We highlight two properties of $\widehat{V}(x,z,w)$. First, unlike traditional series estimators, the regression coefficient $\widehat{\gamma}^L(x)$ in (30) depends on x because the dependent variable in V is a function of x. This fact makes the convergence rate of $\widehat{V}$ slower than the standard rates for series estimators (Imbens and Newey, 2009). Second, a trimming function τ is applied to $q^L(z,w)' \widehat{\gamma}^L(x)$ because we estimate a conditional CDF which by definition lies between zero and one. One example of τ is $\tau(x) = \mathbb{1}\{x \geq 0\} \times \min(x, 1)$.

Next, we estimate $G(s)$ by regressing Y_{it} on the basis functions $p^K(\cdot)$ of $\left(X_{it}, \widehat{V}_{it}, W_i\right)$:

$$\widehat{G}(s) = p^K(s)' \widehat{P}^{-1} \widehat{p}' y/n =: p^K(s)' \widehat{\alpha}^K. \qquad (31)$$

Following Newey, Powell, and Vella (1999), we construct the basis function $p^K(s) = x \otimes p^{K_1}(v,w)$ by exploiting the index structure of the model (1). The index structure enables a faster convergence rate for $\widehat{G}(s)$. Note that in (31) $\widehat{V}_{it}$ from the first-step

is plugged in wherever V_{it} appears.

Finally, we estimate $\beta(v,w)$ by exploiting the index structure of the model (1) and calculate it as

$$\widehat{\beta}(v,w) = \partial\widehat{G}(s)/\partial x = \left(I_{d_X} \otimes p^{K_1}(v,w)\right)' \widehat{\alpha}^K =: \overline{p}(s)' \widehat{\alpha}^K, \tag{32}$$

where the second equality holds by the chain rule. To estimate $\beta(x)$ and $\overline{\beta}$, we use the LIE and regress $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ on the basis function $r^M(\cdot)$ of X_{it} and the constant 1, respectively :

$$\begin{aligned} \widehat{\beta}(x) &= r^M(x)' \widehat{R}^{-1} r' \widehat{B}/n =: r^M(x)' \widehat{\eta}^M, \\ \widehat{\overline{\beta}} &= n^{-1} \sum_{i=1}^{n} \widehat{\beta}\left(\widehat{V}_{it}, W_i\right). \end{aligned} \tag{33}$$

One may consider $\widehat{\overline{\beta}}$ as a "special case" of $\widehat{\beta}(x)$ by letting $r^M(\cdot) \equiv 1$, which simplifies the asymptotic analysis in the next section.

The objects of interest in this paper are $\beta(v,w)$, $\beta(x)$, and $\overline{\beta}$. $\beta(v,w)$ is the conditional expectation of β_{it} given $(V_{it}, W_i) = (v,w)$, and can be interpreted as the average of the partial effects of X_{it} on Y_{it} among the individuals with the same $(V_{it}, W_i) = (v,w)$. If one loosely considers V_{it} to be U_{it} and W_i to be A_i, then $\beta(v,w)$ is the same as β_{it}. In this sense, $\beta(v,w)$ provides the "finest" approximation of β_{it} among the three objects in (29). $\beta(x)$ measures the average partial effect averaged over the conditional distribution of the unobserved heterogeneity (A_i, U_{it}) when X_{it} equals x. It provides useful information about the partial effects of X_{it} on Y_{it} for a subpopulation characterized by $X_{it} = x$. For example, if one asks about the average output elasticity with respect to labor for firms with a certain level of capital and labor, then $\beta(x)$ contains relevant information to answer such questions. $\overline{\beta}$ is the APE that has been studied extensively in the literature (Chamberlain, 1984, 1992; Wooldridge, 2005b; Graham and Powell, 2012; Laage, 2020). It is interpreted as the average of the partial effect of X_{it} on Y_{it} over the unconditional distribution of (A_i, U_{it}). Depending on the scenario and application, all three objects can be useful to answer policy-related questions.

The multi-step series estimators proposed in this section cause challenges for inference due to their multi-layered nature. To obtain large sample properties of $\widehat{\beta}(v,w)$, $\widehat{\beta}(x)$ and $\widehat{\overline{\beta}}$, one needs to analyze the estimators step by step as the estimator from

each step is plugged in and thus affects all subsequent ones. For asymptotic analysis, there is a key difference between $\beta(v,w)$ and $\beta(x)$ or $\overline{\beta}$: $\beta(v,w)$ is a *known* functional of $G(s)$, whereas both $\beta(x)$ and $\overline{\beta}$ are *unknown* but identifiable functionals of $G(s)$. We present in the next session how to deal with these challenges for the purpose of inference.

4.2 Large Sample Theory

We now derive convergence rates and asymptotic normality results for the proposed estimators. Since we let $n \to \infty$ for each t in the asymptotic analysis, the t-subscript is suppressed for notational simplicity. First, we obtain convergence rates for $\widehat{\beta}(v,w)$, $\widehat{\beta}(x)$, and $\widehat{\overline{\beta}}$, respectively. For $\widehat{\beta}(v,w)$, we adapt the results of Imbens and Newey (2009) to our TERC model. For $\widehat{\beta}(x)$ and $\widehat{\overline{\beta}}$, the effects from first- and second- step estimations need to be taken into consideration. We present both mean squared and uniform rates for all three estimators.

Then, we prove asymptotic normality for the estimators, and show that the corresponding variances can be consistently estimated to construct valid confidence intervals. Asymptotic normality for $\widehat{\beta}(v,w)$ is established by applying the results of Andrews (1991) and Imbens and Newey (2002) to cover vector-valued functionals. For $\widehat{\beta}(x)$ and $\widehat{\overline{\beta}}$, the main difference from the existing literature is that both estimators are *unknown* functionals of $G(\cdot)$ that are only estimable from the data. Therefore, one needs to correctly account for the additional estimation error and adjust the asymptotic variance.

Convergence Rates

Recall that the conditional and unconditional moments of the random coefficients are estimated via the three-step estimators (32)–(33). The convergence rates for the first- and second- step estimators $\widehat{V}$ and $\widehat{G}$ have been obtained in Imbens and Newey (2009). We adapt their results to our TERC model and impose the following regularity assumption.

Assumption 7. *Suppose the following conditions hold:*

1. *There exist d_1, $C > 0$ such that for every L there is a $L \times 1$ vector $\gamma^L(x)$*

satisfying

$$\sup_{x\in\mathcal{X},z\in\mathcal{Z},w\in\mathcal{W}}\left|F_{X|Z,W}\left(x|\,z,w\right)-q^{L}\left(z,w\right)'\gamma^{L}\left(x\right)\right|\leq CL^{-d_1}.$$

2. *The joint density of* (X,V,W) *is bounded above and below by constant multiples of its marginal densities.*

3. *There exist* $C>0$, $\zeta\left(K_1\right)$, *and* $\zeta_1\left(K_1\right)$ *such that* $\zeta\left(K_1\right)\leq C\zeta_1\left(K_1\right)$ *and for each* K_1 *there exists a normalization matrix* B *such that* $\widetilde{p}^{K_1}\left(v,w\right)=Bp^{K_1}\left(v,w\right)$ *satisfies* $\lambda_{\min}\left(\mathbb{E}\widetilde{p}^{K_1}\left(V_i,W_i\right)\widetilde{p}^{K_1}\left(V_i,W_i\right)'\right)\geq C$, $\sup_{v\in\mathcal{V},w\in\mathcal{W}}\left\|\widetilde{p}^{K_1}\left(v,w\right)\right\|\leq C\zeta\left(K_1\right)$, *and* $\sup_{v\in\mathcal{V},w\in\mathcal{W}}\left\|\partial\widetilde{p}^{K_1}\left(v,w\right)/\partial v\right\|\leq C\zeta_1\left(K_1\right)$. *Furthermore,* $K_1\zeta_1\left(K_1\right)^2\left(L/n+L^{1-2d_1}\right)$ *is* $o\left(1\right)$.

4. $G\left(s\right)$ *is Lipschitz in* v. *There exist* d_2, $C>0$ *such that for every* $K=d_X\times K_1$ *there is a* $K\times 1$ *vector* α^K *satisfying*

$$\sup_{s\in\mathcal{S}}\left|G\left(s\right)-p^{K}\left(s\right)'\alpha^{K}\right|\leq CK^{-d_2}.$$

5. $Var\left(Y_i|\,X_i,Z_i,W_i\right)$ *is bounded uniformly over the support of* (X_i,Z_i,W_i).

Assumption 7(1) and (4) specify the approximation rates for the series estimators. It is well-known that such rates exist when $F_{X|Z,W}\left(x|\,z,w\right)$ and $G\left(s\right)$ satisfy mild smoothness conditions and regular basis functions like splines are used. See Imbens and Newey (2009) for a detailed discussion.

Assumption 7(2) is imposed to guarantee that the smallest eigenvalue of $\mathbb{E}p^{K}\left(S_i\right)p^{K}\left(S_i\right)'$ is strictly larger than some positive constant C. It is imposed because in the analysis we exploit the index structure of our TERC model by choosing $p^{K}\left(s\right)=x\otimes p^{K_1}\left(v,w\right)$. The usual normalization (Newey, 1997) on the second moment of basis functions can only be done on x and $p^{K_1}\left(v,w\right)$ separately. Thus, we need Assumption 7(2) to make sure the second moment of $p^{K}\left(s\right)$ is well-behaved. A similar assumption is imposed in Imbens and Newey (2002) as well.

Assumption 7(3) is a normalization on the basis function $p^{K_1}\left(\cdot\right)$, which ensures that one can normalize $\mathbb{E}p^{K_1}\left(V_i,W_i\right)p^{K_1}\left(V_i,W_i\right)'$ to be the identity matrix I as in Newey (1997). Finally, the conditional variance of Y given (X,V,W) is assumed to be bounded in Assumption 7(5), which is common in the series estimation literature.

With Assumption 7 in position, we prove the following lemma.

Lemma 3 (**First- and Second-Step Convergence Rates**). *Suppose the conditions of Theorem 1 and Assumption 7 are satisfied. Then, we have*

$$
\begin{aligned}
&n^{-1}\sum_{i}\left(\widehat{V}_i - V_i\right)^2 = O_P\left(L/n + L^{1-2d_1}\right) =: O_P\left(\Delta_{1n}^2\right)\\
&\int\left[\widehat{G}(s) - G(s)\right]^2 dF(s) = O_P\left(K_1/n + K_1^{-2d_2} + \Delta_{1n}^2\right) =: O_P\left(\Delta_{2n}^2\right)\\
&\sup_{s\in\mathcal{S}}\left|\widehat{G}(s) - G(s)\right| = O_P\left(\zeta(K_1)\Delta_{2n}\right).
\end{aligned}
$$

Lemma 3 states that the mean squared convergence rate for $\widehat{G}$ is the sum of the first-step rate Δ_{1n}^2, the variance term K_1/n, and the squared bias term $K_1^{-2d_2}$. Both d_1 and d_2 are the uniform approximation rates that govern how well one is able to approximate the unknown functions V and G with $q^L(\cdot)$ and $p^K(\cdot)$, respectively. Note that even though the order of the basis function for the second-step estimation is K, by the TERC structure $K = d_X \times K_1$ and d_X is a finite constant. Thus, the effective order that matters for the convergence rate results is K_1.

We now obtain the convergence rates for $\widehat{\beta}(v,w)$, $\widehat{\beta}(x)$ and $\widehat{\overline{\beta}}$. We impose the following assumption.

Assumption 8. *Suppose the following conditions hold:*

1. *There exist $d_3, C > 0$ such that for every M there is a $M \times d_X$ matrix η^M satisfying*
$$\sup_{x\in\mathcal{X}}\left\|\beta(x) - r^M(x)'\eta^M\right\| \le CM^{-d_3}.$$

2. *There exist $C > 0$ and $\zeta(M)$ such that for each M there exists a normalization matrix B such that $\widetilde{r}^M(x) = Br^M(x)$ satisfies $\lambda_{\min}\left(\mathbb{E}\widetilde{r}^M(X_i)\widetilde{r}^M(X_i)'\right) \ge C$ and $\sup_{x\in\mathcal{X}}\left\|\widetilde{r}^M(x)\right\| \le C\zeta(M)$.*

3. *Let $\xi_i = \beta(V_i, W_i) - \beta(X_i)$ and $\xi = (\xi_1, ..., \xi_n)'$. Then, $\mathbb{E}\left[\xi\xi'\middle|\mathbf{X}\right] \le CI$ in the positive definite sense.*

4. *$\beta(v,w)$ is Lipschitz in v, with the Lipschitz constant bounded from above.*

Assumption 8 imposes conditions on the approximation rate of $\beta(x)$, the normalization of basis functions $r^M(x)$, and the boundedness of the second moment of ξ_i, similarly to those in Assumption 7.

Theorem 2 (**Third-Step Convergence Rates**). *Suppose the conditions of Lemma 3 and Assumption 8 are satisfied. Then, we have*

$$\begin{aligned}
&\int \left\|\widehat{\beta}(v,w)-\beta(v,w)\right\|^2 dF(v,w) = O_P\left(\Delta_{2n}^2\right),\\
&\int \left\|\widehat{\beta}(x)-\beta(x)\right\|^2 dF(x) = O_P\left(\Delta_{2n}^2 + M/n + M^{-2d_3}\right) =: O_p\left(\Delta_{3n}^2\right),\\
&\left\|\widehat{\overline{\beta}}-\overline{\beta}\right\|^2 = O_P\left(\Delta_{2n}^2\right),\\
&\sup_{v\in\mathcal{V},w\in\mathcal{W}} \left\|\widehat{\beta}(v,w)-\beta(v,w)\right\| = O_P\left(\zeta(K_1)\Delta_{2n}\right), \text{ and}\\
&\sup_{x\in\mathcal{X}} \left\|\widehat{\beta}(x)-\beta(x)\right\| = O_P\left(\zeta(M)\Delta_{3n}\right).
\end{aligned}$$

The first three equations in Theorem 2 give mean squared convergence rates, while the last two show uniform ones. For $\widehat{\beta}(v,w)$, the convergence rate is the same as $\widehat{G}$ because they share the same regression coefficient $\widehat{\alpha}^K$ and only differ in the basis functions used. More precisely, for $\widehat{\beta}(v,w)$ we use $I_{d_X}\otimes p^{K_1}(v,w)$, while for $\widehat{G}(s)$ we use $x\otimes p^{K_1}(v,w)$. Meanwhile, the same regression coefficient $\widehat{\alpha}^K$ is used for both estimators. Therefore, under Assumption 7 and 8, the convergence rate result on $\widehat{G}(s)$ applies directly to $\widehat{\beta}(v,w)$.

For $\widehat{\beta}(x)$ and $\widehat{\overline{\beta}}$, further analysis is required because both estimators involve an additional estimation step. Specifically, for $\widehat{\beta}(x)$, we estimate it with

$$\widehat{\beta}(x) = r^M(x)'\left(\widehat{R}^{-1}r'\widehat{B}/n\right) =: r^M(x)'\widehat{\eta}^M. \tag{34}$$

To obtain the convergence rate for $\widehat{\beta}(x)$, the key steps include expanding

$$\widehat{\eta}^M - \eta^M = \widehat{R}^{-1}r'\left[\left(\widehat{B}-\widetilde{B}\right)+\left(\widetilde{B}-B\right)+\left(B-B^X\right)+\left(B^X-r\eta^M\right)\right]/n, \tag{35}$$

where η^M is defined in Assumption 8(1), and deriving the rate for each component. We show the proof in the Appendix B.

For $\widehat{\overline{\beta}}$, we estimate it with

$$\widehat{\overline{\beta}} = n^{-1} \sum_i \widehat{\beta}\left(\widehat{V}_i, W_i\right). \tag{36}$$

It is possible to analyze $\widehat{\overline{\beta}}$ in a similar way as $\widehat{\beta}(x)$ by expanding $\widehat{\beta}\left(\widehat{V}_i, W_i\right) - \overline{\beta}$ stochastically and deriving the convergence rate component by component. However, with the convergence results established for $\widehat{\beta}(x)$, one can let $r^M(\cdot) \equiv 1$ in (34) and directly obtain the rate for $\widehat{\overline{\beta}}$. We follow this simpler approach in the proof.

Asymptotic Normality

In this section, we prove asymptotic normality for the estimators of $\beta(v, w)$, $\beta(x)$ and $\overline{\beta}$, and show that the corresponding covariance matrices can be consistently estimated for use in confidence intervals. Imbens and Newey (2002) have obtained asymptotic normality for estimators of known and scalar-valued linear functionals of $G(s)$. However, $\beta(v, w)$ is a known but vector-valued functional of $G(s)$. To apply their results, we use Assumption J(iii) of Andrews (1991) together with a Cramér–Wold device to show asymptotic normality for $\widehat{\beta}(v, w)$.

Assumption 9. *Suppose the following conditions hold:*

1. *There exist $C > 0$ and $\zeta(L)$ such that for each L there exists a normalization matrix B such that $\widetilde{q}^L(z, w) = Bq^L(z, w)$ satisfies $\lambda_{\min}\left(\mathbb{E}\widetilde{q}^L(Z_i, W_i)\,\widetilde{q}^L(Z_i, W_i)'\right) \geq C$ and $\sup_{z\in\mathcal{Z}, w\in\mathcal{W}} \left\|\widetilde{q}^L(z, w)\right\| \leq C\zeta(L)$.*

2. *$G(s)$ is twice continuously differentiable with bounded first and second derivatives. For functional $a(\cdot)$ of G and some constant $C > 0$, it is true that $|a(G)| \leq C \sup_s |G(s)|$ and either (i) there is $\delta(s)$ and $\widetilde{\alpha}^K$ such that $\mathbb{E}\delta(S_i)^2 < \infty$, $a\left(p_k^K\right) = \mathbb{E}\delta(S_i)\, p_k^K(S_i)$ for all $k = 1, ..., K$, $a(G) = \mathbb{E}\delta(S_i)\, G(S_i)$, and $\mathbb{E}\left(\delta(S_i) - p^K(S_i)'\,\widetilde{\alpha}^K\right)^2 \to 0$; or (ii) for some $\widetilde{\alpha}^K$, $\mathbb{E}\left[p^K(S_i)'\,\widetilde{\alpha}^K\right]^2 \to 0$ and $a\left(p^K(\cdot)'\,\widetilde{\alpha}^K\right)$ is bounded away from zero as $K \to \infty$.*

3. *$\mathbb{E}\left[(Y - G(s))^4 \middle| X, Z, W\right] < \infty$ and $Var(Y|X, Z, W) > 0$.*

4. *nL^{1-2d_1}, nK^{-2d_2}, $K\zeta_1(K)^2 L^2/n$, $\zeta(K)^6 L^4/n$, $\zeta_1(K)^2 LK^{-2d_2}$, and $\zeta(K)^4 \zeta(L)^4 L/n$ are $o(1)$.*

5. *There exist d_4 and $\overline{\alpha}^K$ such that for each element s_j of $s = (x, v, w)'$:*

$$\max\left\{\sup_{s\in\mathcal{S}}\left|G(s) - p^K(s)'\,\overline{\alpha}^K\right|, \sup_{s\in\mathcal{S}}\left|\partial\left(G(s) - p^K(s)'\,\overline{\alpha}^K\right)/\partial s_j\right|\right\} = O\left(K^{-d_4}\right).$$

6. *(As' J(iii) of Andrews (1991)) For a bounded sequence of constants $\{b_{1n} : n \geq 1\}$ and constant pd matrix $\overline{\Omega}_1$, it is true that $b_{1n}\Omega_1 \xrightarrow{p} \overline{\Omega}_1$.*

Assumptions 9(1)–(5) are imposed in Imbens and Newey (2002) and are regularity conditions required for the asymptotic normality of $\widehat{\beta}(v, w)$. See Newey (1997) for a detailed discussion of these assumptions. Assumption 9(6) is used in Andrews (1991) and guarantees that the normality result of Imbens and Newey (2002) applies to vector-valued functionals of $G(s)$. Essentially, it requires all the coordinates of $\widehat{\beta}(v, w)$ to converge at the same speed, which is a mild assumption under our settings because ex-ante we do not distinguish one coordinate of β_{it} from the others.

Theorem 3 (**Asymptotic Normality for $\widehat{\boldsymbol{\beta}}(\boldsymbol{v}, \boldsymbol{w})$**). *Suppose the conditions of Theorem 2 and Assumption 9 are satisfied. Then, we have*

$$\sqrt{n}\widehat{\Omega}_1^{-1/2}\left(\widehat{\beta}(v, w) - \beta(v, w)\right) \xrightarrow{d} N(0, I).$$

It is worth noting that $\widehat{\Omega}_1$ in Theorem 3 is a function of (v, w), which is omitted for simplicity of exposition. Theorem 3 concerns $\beta(v, w)$, a *known* functional of $G(s)$. However, the result does not directly apply to $\beta(x)$ and $\overline{\beta}$, because they are *unknown* functionals of $G(s)$ and both require an additional estimation step. More specifically, by the LIE one has

$$\beta(x) = \mathbb{E}\left[\partial G(S_i)/\partial X \,|\, X_i = x\right], \ \overline{\beta} = \mathbb{E}\left[\partial G(S_i)/\partial X\right], \tag{37}$$

both of which involve integrating over the unknown but estimable distribution of (V_i, W_i). Therefore, one need estimate these unknown functionals and correctly account for the bias arising from this additional estimation step in asymptotic analysis.

Assumption 10. *Suppose the following conditions hold:*

1. *There exists $C > 0$ such that for each M and K there exist normalization matrices B_1 and B_2 such that $\widetilde{r}^M(x) = B_1 r^M(x)$ and $\widetilde{\overline{p}}^K(s) = B_2\overline{p}^K(s)$ satisfy $\lambda_{\min}\left(\mathbb{E}\widetilde{r}^M(X_i)\,\widetilde{r}^M(X_i)'\right) \geq C$, $\lambda_{\min}\left(\mathbb{E}\widetilde{\overline{p}}^K(S_i)\,\widetilde{\overline{p}}^K(S_i)'\right) \geq$*

C, $\lambda_{\min}\left(\mathbb{E}\widetilde{r}^{M}(X_i)\,\widetilde{\overline{p}}^{K}(S_i)'\left(\mathbb{E}p^{K}(S_i)\,p^{K}(S_i)'\right)^{-1}\mathbb{E}\widetilde{\overline{p}}^{K}(S_i)\,\widetilde{r}^{M}(X_i)'\right) \geq C$, $\sup_{x\in\mathcal{X}}\left\|\widetilde{r}^{M}(x)\right\| \leq C\zeta(M)$, *and* $\sup_{s\in\mathcal{S}}\left\|\widetilde{\overline{p}}^{K}(s)\right\| \leq C\zeta(K)$.

2. *The fourth order moment of* $\xi_i := \beta(V_i, W_i) - \beta(X_i)$ *satisfies* $\mathbb{E}\left[\xi_i^4 | X_i\right] < \infty$.

3. *For a sequence of bounded constants* $\{b_{2n} : n \geq 1\}$ *and some constant pd matrix* $\overline{\Omega}_2$, $b_{2n}\Omega_2 \xrightarrow{p} \overline{\Omega}_2$ *holds.*

Assumption 10(1) is a normalization on basis functions $r^M(\cdot)$ and $\overline{p}^K(\cdot)$. The substantial part is

$$\lambda_{\min}\left(\mathbb{E}\widetilde{r}^{M}(X_i)\,\widetilde{\overline{p}}^{K}(S_i)'\left(\mathbb{E}p^{K}(S_i)\,p^{K}(S_i)'\right)^{-1}\mathbb{E}\widetilde{\overline{p}}^{K}(S_i)\,\widetilde{r}^{M}(X_i)'\right) \geq C, \tag{38}$$

which is needed to show that the asymptotic covariance matrix Ω_2 of $\sqrt{n}\left(\widehat{\beta}(x) - \beta(x)\right)$ is positive definite. Assumption 10(2) is a regularity condition imposed for the Lindeberg–Feller Central Limit Theorem (CLT). Assumption 10(3) is similar to Assumption 9(6) and is needed to show the asymptotic normality result holds for vector-valued functionals of $G(s)$.

Theorem 4 (**Asymptotic Normality for $\widehat{\boldsymbol{\beta}}(\boldsymbol{x})$ and $\widehat{\overline{\boldsymbol{\beta}}}$**). *Suppose the conditions of Theorem 3 and Assumption 10 are satisfied. Then, we have*

$$\sqrt{n}\widehat{\Omega}_2^{-1/2}\left(\widehat{\beta}(x) - \beta(x)\right) \xrightarrow{d} N(0, I).$$

Furthermore, if $\mathbb{E}\left\|\beta(v,w) - \overline{\beta}\right\|^4 < \infty$ *, we have*

$$\sqrt{n}\widehat{\Omega}_3^{-1/2}\left(\widehat{\overline{\beta}} - \overline{\beta}\right) \xrightarrow{d} N(0, I).$$

Theorem 4 gives the asymptotic normality results that can be used to construct confidence intervals and test statistics for both $\beta(x)$ and $\overline{\beta}$. To see why the results of Imbens and Newey (2002) are not directly applicable, suppose β is a scalar and let $\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) := \widehat{\beta}(x)$ and $a(\beta, V) := \beta(x)$. Then, we have

$$\begin{aligned} &\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - a(\beta, V) \\ &= \underbrace{\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - \widehat{a}\left(\beta, \widehat{V}\right)}_{\text{known functional of } G(s)} + \underbrace{\widehat{a}\left(\beta, \widehat{V}\right) - \widehat{a}(\beta, V)}_{\text{estimation error of } V} + \underbrace{\widehat{a}(\beta, V) - a(\beta, V)}_{\text{estimation erro of } a}. \end{aligned} \tag{39}$$

From (39), it is clear that because one needs to estimate both unknown functional a and unknown random variable V, in addition to the first term in (39) that concerns a known functional of $G(s)$, there are two more terms that affects the asymptotic normality of $\beta(x)$. In Appendix B, we show how to correctly account for the effects from both estimation steps on influence functions. It is worth mentioning that for $\widehat{\overline{\beta}}$ one can significantly simplify the analysis by observing that $\widehat{\overline{\beta}}$ can be viewed as a "special case" of $\widehat{\beta}(x)$, that is, choosing $r^M(\cdot) \equiv 1$ in the definition of $\widehat{\beta}(x)$ gives $\widehat{\overline{\beta}}$. Therefore, with slight modifications to the proof for $\widehat{\beta}(x)$ one proves normality for $\widehat{\overline{\beta}}$.

5 Simulation

In this section, we examine the finite-sample performance of the method via a simulation study. A discussion of the dgp motivated by production function applications is first provided. Then, we show the baseline results and compare the distribution of the estimated random coefficients with the simulated ones. In the Appendix, we run several robustness checks to investigate how our method performs when one varies the number of periods and firms, as well as the orders of the basis functions used for the series estimation, and when one includes exogenous shocks to the dgp.

5.1 DGP

The baseline dgp we consider is

$$Y_{it} = X_{it}^K \beta_{it}^K + X_{it}^L \beta_{it}^L + \omega_{it}, \tag{40}$$

where the random coefficients $\left(\omega_{it}, \beta_{it}^K, \beta_{it}^L\right)$ are functions of (A_i, U_{it}), X_{it}^K and X_{it}^L are input choices of (natural log of) capital and labor, and Y_{it} is the (natural log of) output. To allow correlation between A_i and U_{it}, an important feature in empirical applications, we draw $A_i \sim \mathcal{U}[1,2]$ and let $U_{it} = A_i \times \eta_{it}^I + \eta_t^{II}$ where $\eta_{it}^I \sim \mathcal{U}[1,3/2]$ and $\eta_t^{II} \sim \mathcal{U}[1,3/2]$ capture idiosyncratic and macro shocks, respectively. Then, we construct the random coefficients as $\omega_{it} = U_{it}$, $\beta_{it}^K = A_i + U_{it}$, and $\beta_{it}^L = A_i \times U_{it}$ and let $\beta_{it} = \left(\omega_{it}, \beta_{it}^K, \beta_{it}^L\right)'$. Thus, we have a total of $N \times T \times B$ β_{it}'s where N, T and B are total number of firms, periods, and simulations, respectively. Based on the DGP, we calculate the true $\overline{\omega} := \mathbb{E}\omega_{it} = 25/8$ and APEs of $\overline{\beta}^K := \mathbb{E}\beta_{it}^K = 37/8$ and $\overline{\beta}^L := \mathbb{E}\beta_{it}^L = 115/24$. We further define $\overline{\beta} := \left(\overline{\omega}, \overline{\beta}^K, \overline{\beta}^L\right)'$. Finally, we draw

each element of the instrument $Z_{it} = (R_{it}, W_{it}, P_{it})'$ independently from $\mathcal{U}[1,3]$, and calculate capital X_{it}^K and labor X_{it}^L by solving a representative firm's static profit maximization problem

$$X_{it}^K = \left[\left(1-\beta_{it}^L\right)\ln\left(R_{it}/\beta_{it}^K\right) + \beta_{it}^L \ln\left(W_{it}/\beta_{it}^L\right) - \ln\left(\omega_{it}P_{it}\right)\right] / \left(\beta_{it}^K + \beta_{it}^L - 1\right),$$
$$X_{it}^L = \left[\left(1-\beta_{it}^K\right)\ln\left(W_{it}/\beta_{it}^L\right) + \beta_{it}^K \ln\left(R_{it}/\beta_{it}^K\right) - \ln\left(\omega_{it}P_{it}\right)\right] / \left(\beta_{it}^K + \beta_{it}^L - 1\right).$$

Note that we do not include the ex-post shocks ε_{it} nor the exogenous productivity shock $U_{2,it}$ for the baseline scenario, but will add them later on to investigate how they affect the performance.

In the simulations, the observable data are (X, Y, Z), which are used to estimate $\beta(v,w)$, $\beta(x)$, and $\overline{\beta}$ via the three-step estimation outlined in Section 4.1. Then, we evaluate the performance of the estimated $\widehat{\beta}(v,w)$, $\widehat{\beta}(x)$, and $\widehat{\overline{\beta}}$ against the truth.

5.2 Baseline Results

For the baseline configuration, we set $N = 1000$ and $T = 3$, and use basis functions of degree two splines with knot at the median. We run $B = 100$ simulations and summarize the performance of $\widehat{\overline{\omega}}, \widehat{\overline{\beta}}^K$ and $\widehat{\overline{\beta}}^L$ in Table 1.

Table 1: Performance of $\widehat{\overline{\omega}}, \widehat{\overline{\beta}}^K$ and $\widehat{\overline{\beta}}^L$

	Formula	$\widehat{\overline{\omega}}$	$\widehat{\overline{\beta}}^K$	$\widehat{\overline{\beta}}^L$
Bias	$B^{-1}\sum_b\left(\widehat{\overline{\beta}}_b^{(d)} - \overline{\beta}^{(d)}\right) / \left\lvert\overline{\beta}^{(d)}\right\rvert$	0.0119	0.0144	0.0066
rMSE	$\sqrt{B^{-1}\sum_b\left(\widehat{\overline{\beta}}_b^{(d)} - \overline{\beta}^{(d)}\right)^2} / \left\lvert\overline{\beta}^{(d)}\right\rvert$	0.0318	0.0257	0.0323

Table 1 shows that the proposed method can accurately estimate the APE $\overline{\beta}$. Specifically, the first row evaluates the performance based on the normalized average bias for each coordinate of $\overline{\beta}$ across B rounds of simulations. The bias is small for all three coordinates, with a magnitude between 0.66% and 1.44% of the length of corresponding $\overline{\beta}^{(d)}$. The second row measures the normalized rMSE of $\widehat{\overline{\beta}}$ against true $\overline{\beta}$ for each coordinate, and shows that the method is able to achieve a low rMSE between 2.57% and 3.23% of the length of corresponding $\overline{\beta}^{(d)}$. By the standard

bias-variance decomposition of MSE, the results in Table 1 show that the bias of the estimator for the APE is dominated by its variance.

Next, since $\beta(V_{it}, W_i)$ can be thought of as the "finest" approximation of β_{it}, one may wonder how closely the distribution of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ mimics that of true β_{it}. In the following analysis, we compare the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ with that of true β_{it} to show how accurately the method can capture the distributional properties of the random coefficients.

Figure 2–4 show the histogram of each coordinate of the estimated (brown) $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ versus that of true (blue) β_{it}. In all three figures, the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ centers around the corresponding population mean. It is worth mentioning that the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ seems more centered around its mean with slightly thinner tails than the corresponding coordinate of the simulated β_{it}. The phenomenon is possibly caused by the fact that $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ is an estimator of $\mathbb{E}\left[\beta_{it} | V_{it}, W_i\right]$ and thus already involves some averaging. Nonetheless, it is evident in Figure 2–4 that there is significant overlap between the distribution of each coordinate of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ and that of β_{it}, implying that the proposed method can accurately estimate both the mean and the dispersion of the random coefficients.

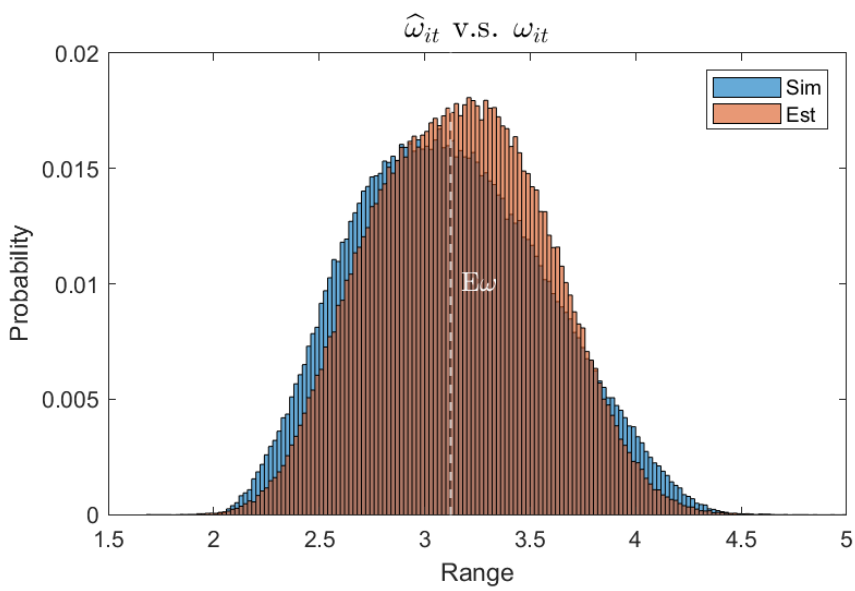

Figure 2: Histogram of $\widehat{\omega}_{it}$ versus ω_{it}

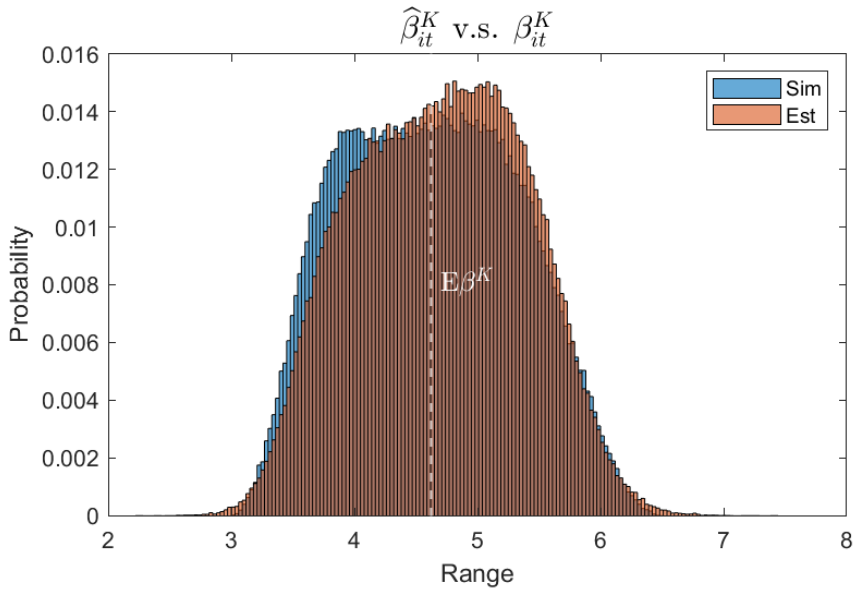

Figure 3: Histogram of $\widehat{\beta}_{it}^K$ versus β_{it}^K

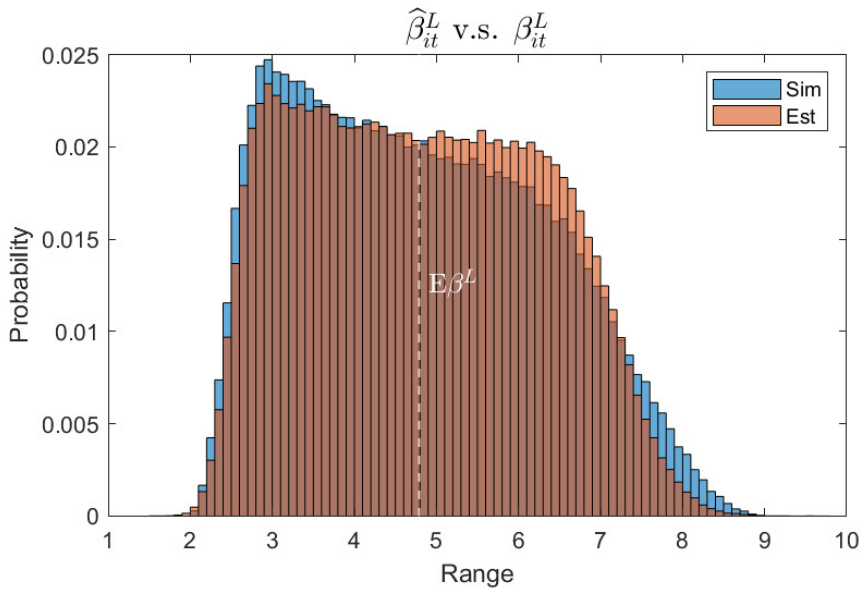

Figure 4: Histogram of $\widehat{\beta}_{it}^L$ versus β_{it}^L

Figure 4 is especially interesting because the true β_{it}^L follows a non-standard distribution that is right-skewed. Still, the histogram of $\widehat{\beta}_{it}^L$ looks very similar to the distribution of β_{it}^L, providing further evidence that the method works well even when the underlying distribution is non-standard.

6 Production Function Application

In this section, we apply the procedure to comprehensive production data for Chinese manufacturing firms. Specifically, we estimate a valued-added production function for each of the five largest sectors in terms of the number of firms, where the output elasticities are allowed to vary across firms and periods, and, more importantly, input choices are allowed to depend on time-varying coefficients in each period in a nonseparable way.

Output elasticity is an essential object of interest in the study of production functions as it quantifies how output responds to variations of each input, e.g., labor, capital, or material. It also helps answer important policy-related questions such as what returns to scale faced by a firm are, how the adoption of a new technology affects production, how the allocation of firm inputs relates to productivity, among others. We estimate conditional and unconditional means of the output elasticities for each sector and compare the results with those obtained using the method of Ackerberg, Caves, and Frazer (2015) (ACF) on the same dataset. Results show that there is substantial variation in the output elasticities in both dimensions, highlighting the importance of properly accounting for both unobserved heterogeneity and time-varying endogeneity.

6.1 Data and Methodology

We use China Annual Survey of Industrial Firms (CASIF), a longitudinal micro-level data collected by the National Bureau of Statistics of China that include information for all state-owned industrial firms and non-state-owned firms with annual sales above 5 million RMB (~US$770K). According to Brandt et al. (2017), they account for 91 percent of the gross output, 71 percent of employment, 97 percent of exports, and 91 percent of total fixed assets in 2004, thus is representative of industrial activities in China. Many research on topics such as firm behavior, international trade, foreign direct investment, and growth theory use the CASIF data. See, for example, Hsieh and Klenow (2009), Song et al. (2011), Brandt et al. (2017), and Roberts et al. (2018).

We focus on the five largest 2-digit sectors [12] in terms of the number of firms for 2004 and 2007 [13]. Following Brandt et al. (2014), appropriate price deflators for inputs

[12] Textile, Chemical, Nonmetallic Minerals, General Equipment, Transportation Equipment.

[13] The CASIF data span between 1998 and 2007. We choose year 2004 and 2007 because there is

and outputs are applied separately. We preprocess data so that firms with strictly positive amount of capital, employment, value-added output, real wage expense and real interests are used for estimation. There are other sanity checks such as total assets should be no smaller than current assets. See Nie et al. (2012) for a detailed discussion.

The final data is a balanced panel with 11,567 firms. The summary statistics of the key variables are presented in Table 2.

Table 2: Summary Statistics

Variables	N	mean	sd	min	max
$y_{it} = \ln(\text{value-added output})$	23,134	9.5287	1.3782	2.4692	16.9649
$k_{it} = \ln(\text{capital})$	23,134	9.1714	1.5908	0.9817	16.8355
$l_{it} = \ln(\text{labor})$	23,134	5.0876	1.0860	2.0794	11.9722
$r_{it} = \ln(\text{real interest rate})$	23,134	0.5234	1.0973	-6.4039	4.5737
$w_{it} = \ln(\text{real wage})$	23,134	2.5482	0.5273	-0.2783	6.1155
Year	2	-	-	2004	2007
Firm ID	11,567	-	-	-	-
Industry Code	5	-	-	-	-

The value-added production function is

$$\begin{aligned}
&y_{it} = \beta_{it}^K k_{it} + \beta_{it}^L l_{it} + \omega_{it} + \varepsilon_{it}, \\
&\beta_{it}^K = \beta^K\left(A_i, U_{it}, U_{2,it}\right), \ \beta_{it}^L = \beta^L\left(A_i, U_{it}, U_{2,it}\right), \ \omega_{it} = \omega\left(A_i, U_{it}, U_{2,it}\right), \\
&k_{it} = g^K\left(Z_{it}, A_i, U_{it}\right), \ l_{it} = g^L\left(Z_{it}, A_i, U_{it}\right), \ X_{it} = \left(k_{it}, l_{it}\right)', \ Z_{it} = \left(r_{it}, w_{it}\right)', \\
&W_i = \text{mean through time of second order polynomials of } \left(X_{it}, Z_{it}\right). \qquad (41)
\end{aligned}$$

We incorporate two features in (41). First, the output elasticities β_{it}^K and β_{it}^L are both allowed to be time-varying and different across firms. Second, model (41) allows input choices X_{it} to depend on $\beta_{it} := \left(\beta_{it}^K, \beta_{it}^L\right)$ in each period via (A_i, U_{it}).

It is worth noting that the output measure is the total revenue in dollars, not physical quantities in pieces due to lack of individual output prices in the data. When

a change in the Chinese Industry Classification codes in 2003 and we intend to use the most recent data with a reasonable gap in time to show the change in the distribution of output elasticities.

firms operate in distinct imperfectly competitive output markets, this may cause issues as pointed out by Klette and Griliches (1996). We use input prices (r_{it}, w_{it}) as instruments Z_{it} for X_{it}[14]. There are other possible choices of instruments including local minimum wage, lagged inputs (De Loecker and Warzynski, 2012; Shenoy, 2020), demand instruments (Goldberg et al., 2010), and product/firm characteristics of direct competitors within the same sector and location (Berry et al., 1995).

We estimate both conditional and unconditional expectations of β_{it} within each two-digit sector. Specifically, we first estimate $V_{it} := F_{k_{it}|Z_{it},W_i}(k_{it}|\, Z_{it}, W_i)$ using second-order splines with its knot at $(10\%, 25\%, 50\%, 75\%, 90\%)$. The choice of the order of basis functions is motivated by simulation results in Section 5. Note that we use k_{it} in V_{it} because it can be considered as a continuous variable while l_{it} is by definition discrete. Next, we estimate $G_{it} := \mathbb{E}[y_{it}|\, X_{it}, V_{it}, W_i]$ using second-order splines again where $\widehat{V}_{it}$ from the first step is plugged in. Finally, we estimate $\beta(V_{it}, W_i) := \mathbb{E}[\beta_{it}|\, V_{it}, W_i]$ by taking the partial derivative of G_{it} with respect to X_{it} and $\overline{\beta} := \mathbb{E}\beta_{it}$ by averaging the estimated $\widehat{\beta}(V_{it}, W_i)$ over i and t.

6.2 Results

Average Elasticities

First, we calculate the average $\widehat{\overline{\beta}} := \left(\widehat{\overline{\beta}}^K, \widehat{\overline{\beta}}^L\right)$ for each sector and compare it with those obtained using ACF's method. The results are summarized in Figure 5.

[14]Real wages are likely to be exogenous because the labor market in the manufacturing sectors of China is close to perfectly competitive where firms face a relatively flat supply curve. Real interest rate may be considered as a valid instrument because its fluctuation is mostly driven by monetary policy set by PBOC.

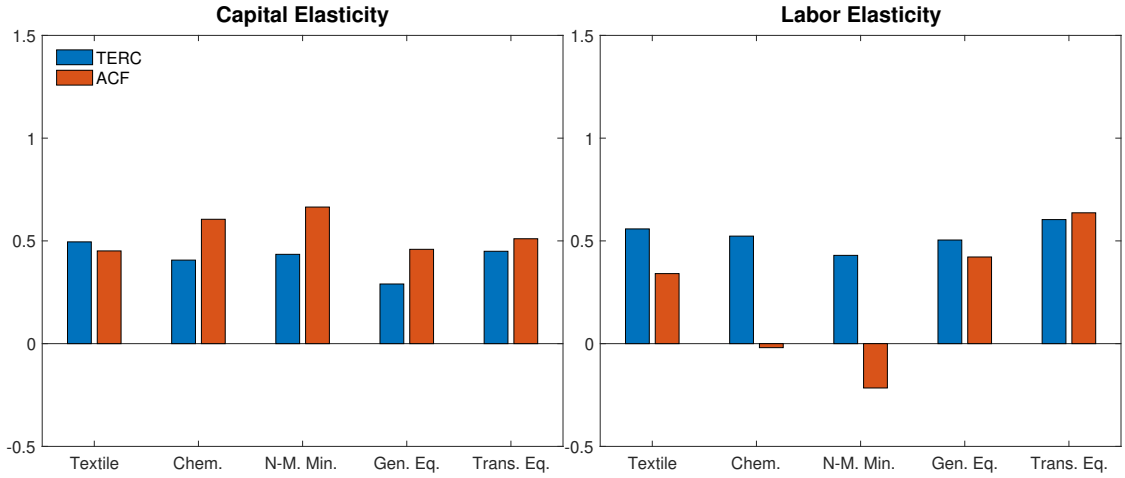

Figure 5: Comparison of Average Elasticities

The blue bars in Figure 5 are the average $\widehat{\overline{\beta}}$ for each sector derived using our method, while the red ones correspond to those obtained using ACF's method. The average capital elasticities $\widehat{\overline{\beta}}^K$ plotted on the left are similar between the two methods, with a magnitude between 0.3 and 0.6. The average labor elasticities $\widehat{\overline{\beta}}^L$ plotted on the right, however, show a significant difference between the two methods. We estimate $\widehat{\overline{\beta}}^L$ to be around 0.5 for each of the five sectors, while applying the method of ACF to the same data gives negative $\widehat{\overline{\beta}}^L$ for two sectors: Chemical and Nonmetallic Minerals, which is hard to rationalize. The application shows that our method is able to generate economically intuitive estimates.

Distribution of Output Elasticities

The distribution of β_{it} may be of interest to policy makers because it provides a more detailed picture of the capital/labor efficiency among the firms. To that end, we investigate the distribution of $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$ within each sector for each year. First, we plot the histogram of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ and $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$ for the Textile sector in Figure 6. Since ACF assume β^K and β^L to be constants, we indicate their location on the same graph.

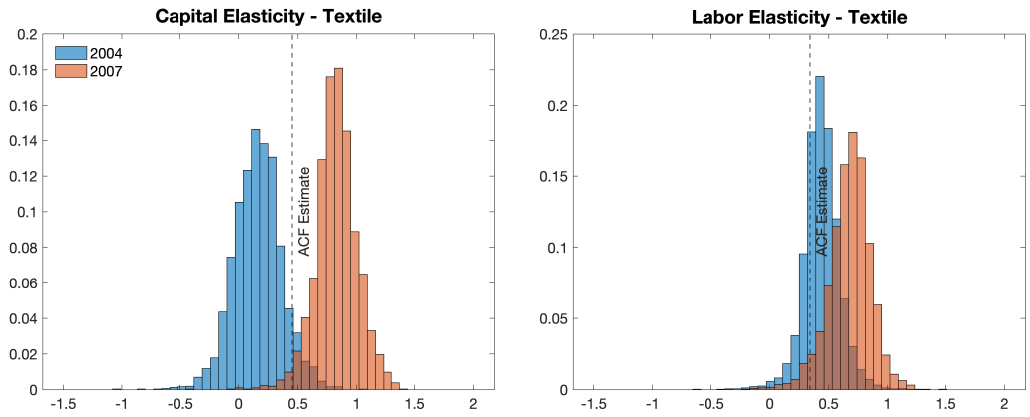

Figure 6: Distribution of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ and $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$: Textile

The left subplot of Figure 6 shows the distribution of capital elasticities for 2004 (blue) and 2007 (orange) in Textile sector. There is a shift in means of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ between the two years, implying an overall improvement in capital efficiency. The majority of the probability mass of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ are located between 0 and 1, which we consider to be economically sensible. We also find mild change in the dispersion of the capital elasticities between 2004 and 2007, with the latter observing slightly more concentrated $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$.

The right subplot of Figure 6 illustrates the distribution of labor elasticity. The change in the mean of $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$ between 2004 and 2007 is not as significant as $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$, suggesting a relatively mild improvement in labor efficiency for the Textile sector. Once again, most of the probability mass of $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$ lies on $[0, 1]$. Lastly, although the estimates of $\overline{\beta}$'s for the Textile sector are close between our method and ACF's method in Figure 5, our method can further allow heterogeneous β_{it} among firms and through time, thus providing more detailed information about the distribution of β_{it}.

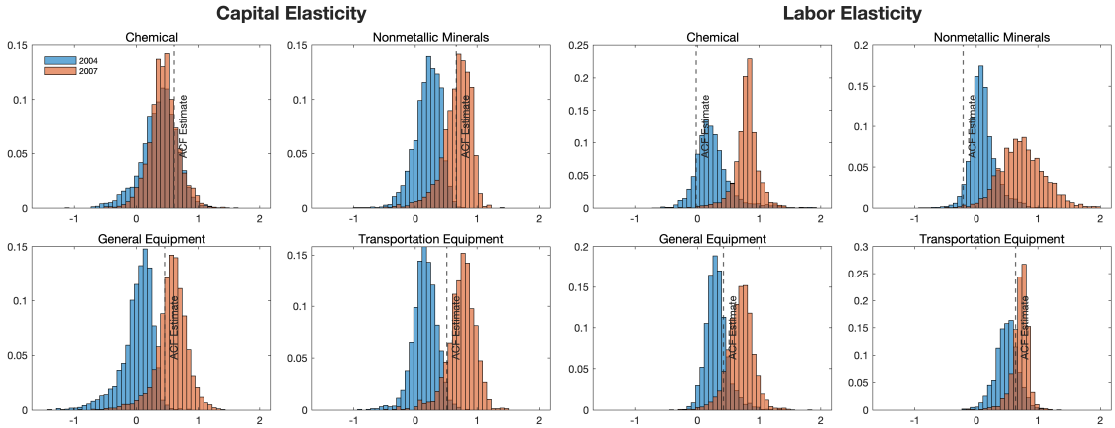

Figure 7: Distribution of $\widehat{\beta}^K\left(\widehat{V}_{it}, W_i\right)$ and $\widehat{\beta}^L\left(\widehat{V}_{it}, W_i\right)$: Other Sectors

In Figure 7, we compare the distribution of capital and labor elasticities between 2004 and 2007 for the other four sectors. Among all sectors, there are improvements in both capital and labor efficiency in terms of the means between the two years. The magnitudes of the changes, however, are more diverse. For example, for the Chemical sector the change in average capital elasticity is close to zero, whereas in the Transportation Equipment sector the shift in means of the capital elasticities is more significant. In the right subplot we observe increase and decrease in the dispersion of labor elasticities for the Nonmetallic Minerals sector and the Transportation Equipment sector, respectively. The findings can be useful for policy makers to adjust industry policies. Finally, we plot the distribution of $\widehat{\beta}\left(X_{it}\right) := \widehat{\mathbb{E}}\left[\beta_{it} \middle| X_{it}\right]$ for each sector in the Appendix E. Not surprisingly, we observe more concentrated histogram due to the additional averaging effect.

7 Conclusion

This paper proposes a flexible random coefficients panel model where the regressors are allowed to depend on the time-varying random coefficients in each period, a critical feature in many economic applications such as production function estimation. The model allows for a nonseparable first-step equation, a nonlinear fixed effect of arbitrary dimension, and an idiosyncratic shock that can be arbitrarily correlated with the fixed effect and that affects the choice of the regressors in a nonlinear way. A sufficiency argument is used to control for the fixed effect, which enables one to

construct a feasible control function for the random shock and subsequently identify the moments of the random coefficients. We provide consistent series estimators for the moments of the random coefficients and prove a new asymptotic normality result. Results from the empirical application highlight the importance of accounting for the potential unobserved heterogeneity and time-varying endogeneity in the data.

We mention several extensions for future research. First, we choose to focus on the first-order moments in this paper. However, one may also be interested in the quantiles of the random coefficients, which may require identifying the pdf of the random coefficients. Second, it is an open question as to whether our method could be extended to dynamic panel data models. Third, given the difficulty of finding appropriate instruments in the applied works, one may want to add timing assumptions (e.g., when agent i chooses its X_{it}) to enable more valid instruments.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification properties of recent production function estimators," *Econometrica*, 83, 2411–2451.

Altonji, J. G. and R. L. Matzkin (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73, 1053–1102.

Andrews, D. W. K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307–45.

Arellano, M. and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models," *The Review of Economic Studies*, 79, 987–1020.

Balestra, P. and M. Nerlove (1966): "Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas," *Econometrica: Journal of the econometric society*, 585–612.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile prices in market equilibrium," *Econometrica: Journal of the Econometric Society*, 841–890.

Blundell, R., X. Chen, and D. Kristensen (2007a): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75, 1613–1669.

Blundell, R., T. MaCurdy, and C. Meghir (2007b): "Labor supply models: Unobserved heterogeneity, nonparticipation and dynamics," *Handbook of econometrics*, 6, 4667–4775.

Brandt, L., J. Van Biesebroeck, L. Wang, and Y. Zhang (2017): "WTO accession and performance of Chinese manufacturing firms," *American Economic Review*, 107, 2784–2820.

Brandt, L., J. Van Biesebroeck, and Y. Zhang (2014): "Challenges of working with the Chinese NBS firm-level data," *China Economic Review*, 30, 339–352.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2012): "Robust inference with multiway clustering," *Journal of Business & Economic Statistics*.

Cameron, A. C. and D. L. Miller (2015): "A practitioner's guide to cluster-robust inference," *Journal of human resources*, 50, 317–372.

Chamberlain, G. (1984): "Panel data," *Handbook of econometrics*, 2, 1247–1318.

——— (1992): "Efficiency bounds for semiparametric regression," *Econometrica: Journal of the Econometric Society*, 567–596.

Chernozhukov, V., J. A. Hausman, and W. K. Newey (2019): "Demand analysis with many prices," Tech. rep., National Bureau of Economic Research.

De Loecker, J. and F. Warzynski (2012): "Markups and firm-level export status," *American economic review*, 102, 2437–71.

Demirer, M. (2020): "Production function estimation with factor-augmenting technology: An application to markups," Tech. rep., MIT working paper.

D'Haultfœuille, X. and P. Février (2015): "Identification of nonseparable triangular models with discrete instruments," *Econometrica*, 83, 1199–1210.

Florens, J.-P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191–1206.

Fox, J. T., V. Haddad, S. Hoderlein, A. K. Petrin, and R. P. Sherman (2016): "Heterogeneous production functions, panel data, and productivity dispersion," *Slide deck, Rice Univ.*

Gandhi, A., S. Navarro, and D. A. Rivers (2020): "On the identification of gross output production functions," *Journal of Political Economy*, 128, 2973–3016.

Goldberg, P. K., A. K. Khandelwal, N. Pavcnik, and P. Topalova (2010): "Imported intermediate inputs and domestic product growth: Evidence from India," *The Quarterly journal of economics*, 125, 1727–1767.

Graham, B. S., J. Hahn, A. Poirier, and J. L. Powell (2018): "A quantile correlated random coefficients panel data model," *Journal of Econometrics*, 206, 305–335.

Graham, B. S. and J. L. Powell (2012): "Identification and estimation of average partial effects in irregular correlated random coefficient panel data models," *Econometrica*, 80, 2105–2152.

Hahn, J. and G. Ridder (2013): "Asymptotic variance of semiparametric estimators with generated regressors," *Econometrica*, 81, 315–340.

——— (2019): "Three-stage semi-parametric inference: Control variables and differentiability," *Journal of econometrics*, 211, 262–293.

Hsiao, C. (2014): *Variable-Coefficient Models*, Cambridge University Press, 167–229, Econometric Society Monographs, 3 ed.

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 124, 1403–1448.

Imbens, G. W. and W. K. Newey (2002): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *NBER Working Paper*.

——— (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481–1512.

Jonnson, W. (1924): *Logic, Part III: Logical Foundation of Science*, Cambridge University Press.

Kasahara, H., P. Schrimpf, and M. Suzuki (2015): "Identification and estimation of production function with unobserved heterogeneity," *University of British Columbia mimeo*.

Kasy, M. (2011): "Identification in triangular systems using control functions," *Econometric Theory*, 27, 663–671.

Klette, T. J. and Z. Griliches (1996): "The inconsistency of common scale estimators when output prices are unobserved and endogenous," *Journal of applied econometrics*, 11, 343–361.

Kyriazidou, E. (1997): "Estimation of a panel data sample selection model," *Econometrica: Journal of the Econometric Society*, 1335–1364.

Laage, L. (2020): "A correlated random coefficient panel model with time-varying endogeneity," *arXiv preprint arXiv:2003.09367*.

Lee, Y.-Y. (2018): "Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models," *Available at SSRN 3250485*.

León-Ledesma, M. A., P. McAdam, and A. Willman (2010): "Identifying the elasticity of substitution with biased technical change," *American Economic Review*, 100, 1330–57.

Levinsohn, J. and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70, 317–341.

Li, T. and Y. Sasaki (2017): "Constructive Identification of Heterogeneous Elasticities in the Cobb-Douglas Production Function," *arXiv preprint arXiv:1711.10031*.

Marschak, J. and W. H. Andrews (1944): "Random simultaneous equations and the theory of production," *Econometrica, Journal of the Econometric Society*, 143–205.

McCall, J. J. (1991): "Exchangeability and its economic applications," *Journal of Economic Dynamics and Control*, 15, 549 – 568.

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica: journal of the Econometric Society*, 69–85.

Newey, W. K. (1994): "The asymptotic variance of semiparametric estimators," *Econometrica: Journal of the Econometric Society*, 1349–1382.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79, 147–168.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric estimation of triangular simultaneous equations models," *Econometrica*, 67, 565–603.

Nie, H., T. Jiang, and R. Yang (2012): "The Current Status and Potential Issues of Chinese Industrial Enterprises Database," *The Journal of World Economy*, 5.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297.

Roberts, M. J., D. Yi Xu, X. Fan, and S. Zhang (2018): "The role of firm factors in demand, cost, and export market selection for Chinese footwear producers," *The Review of Economic Studies*, 85, 2429–2461.

Shenoy, A. (2020): "Estimating the Production Function Under Input Market Frictions," *Review of Economics and Statistics*, 1–45.

Song, Z., K. Storesletten, and F. Zilibotti (2011): "Growing like china," *American economic review*, 101, 196–233.

Stoyanov, J. (2000): "Krein condition in probabilistic moment problems," *Bernoulli*, 6, 939–949.

Torgovitsky, A. (2015): "Identification of nonseparable models using instruments with small support," *Econometrica*, 83, 1185–1197.

Weyl, H. (1939): *The Classical Groups: Their Invariants and Representations.*, Princeton University Press.

Wooldridge, J. M. (2005a): "Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models," *Review of Economics and Statistics*, 87, 385–390.

——— (2005b): "Unobserved heterogeneity and estimation of average partial effects," *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, 27–55.

Appendix

A Proofs in Section 3

Proof of Lemma 1. The proof is divided into two parts. First, we establish the exchangeability condition (11) using Assumption 2. Then, we show that there exist W_i such that (10) holds. For simplicity of notations, we assume X_{it} and Z_{it} are both scalars. The proof goes through when X_{it} and Z_{it} are vectors. We prove (11) for $T = 2$, which is wlog because T is finite and thus any permutation of $(1, ..., T)$ can be achieved by switching pairs of (t_i, t_j) finite number of times. For example, one can obtain (t_3, t_1, t_2) from (t_1, t_2, t_3) by $(t_1, t_2, t_3) \to (t_1, t_3, t_2) \to (t_3, t_1, t_2)$. We suppress i subscripts in all variables in this proof.

By Assumption 2, we have

$$f_{U_1,U_2|A}\left(u_1, u_2 \middle| a\right) = f_{U_1,U_2|A}\left(u_2, u_1 \middle| a\right), \tag{42}$$

which implies

$$f_{A,U_1,U_2}\left(a, u_1, u_2\right) = f_{A,U_1,U_2}\left(a, u_2, u_1\right). \tag{43}$$

Let $g^{-1}\left(X, Z, A\right)$ denote the inverse function of $g\left(Z, A, U\right)$ with respect to U. Define $u_1 = g^{-1}\left(x_1, z_1, a\right)$ and $u_2 = g^{-1}\left(x_2, z_2, a\right)$. Calculate the determinants of the Jacobians as

$$J_1 := \begin{vmatrix} \dfrac{\partial A}{\partial X_1} & \dfrac{\partial A}{\partial X_2} & \dfrac{\partial A}{\partial A} \\ \dfrac{\partial g^{-1}\left(X_1, Z_1, A\right)}{\partial X_1} & \dfrac{\partial g^{-1}\left(X_1, Z_1, A\right)}{\partial X_2} & \dfrac{\partial g^{-1}\left(X_1, Z_1, A\right)}{\partial A} \\ \dfrac{\partial g^{-1}\left(X_2, Z_2, A\right)}{\partial X_1} & \dfrac{\partial g^{-1}\left(X_2, Z_2, A\right)}{\partial X_2} & \dfrac{\partial g^{-1}\left(X_2, Z_2, A\right)}{\partial A} \end{vmatrix}_{\substack{(X_1, X_2, Z_1, Z_2, A) \\ = (x_1, x_2, z_1, z_2, a)}}$$

$$
= \left| \begin{array}{ccc} 0 & 0 & 1 \\ \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial X_1} & 0 & \dfrac{\partial g^{-1}(X_1, Z_1, A)}{\partial A} \\ 0 & \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial X_2} & \dfrac{\partial g^{-1}(X_2, Z_2, A)}{\partial A} \end{array} \right|_{\substack{(X_1, X_2, Z_1, Z_2, A) \\ = (x_1, x_2, z_1, z_2, a)}}
$$

$$
= \partial g^{-1}(X, Z, A) / \partial X \big|_{(X,Z,A)=(x_1,z_1,a)} \times \partial g^{-1}(X, Z, A) / \partial X \big|_{(X,Z,A)=(x_2,z_2,a)}, \tag{44}
$$

and

$$
\begin{aligned}
& J_2 \\
:= & \left| \begin{array}{ccc} \dfrac{\partial g(Z_1, A, U_1)}{\partial A} & \dfrac{\partial g(Z_1, A, U_1)}{\partial U_1} & \dfrac{\partial g(Z_1, A, U_1)}{\partial U_2} \\ \dfrac{\partial g(Z_2, A, U_2)}{\partial A} & \dfrac{\partial g(Z_2, A, U_2)}{\partial U_1} & \dfrac{\partial g(Z_2, A, U_2)}{\partial U_2} \\ \dfrac{\partial A}{\partial A} & \dfrac{\partial A}{\partial U_1} & \dfrac{\partial A}{\partial U_2} \end{array} \right|_{\substack{(Z_1, Z_2, A, U_1, U_2) \\ = (z_2, z_1, a, u_2, u_1)}} \\
= & \left| \begin{array}{ccc} \dfrac{\partial g(Z_1, A, U_1)}{\partial A} & \dfrac{\partial g(Z_1, A, U_1)}{\partial U_1} & 0 \\ \dfrac{\partial g(Z_2, A, U_2)}{\partial A} & 0 & \dfrac{\partial g(Z_2, A, U_2)}{\partial U_2} \\ 1 & 0 & 0 \end{array} \right|_{\substack{(Z_1, Z_2, A, U_1, U_2) \\ = (z_2, z_1, a, u_2, u_1)}} \\
= & \, \partial g(Z, A, U) / \partial U \big|_{(Z,A,U)=(z_2,a,u_2)} \times \partial g(Z, A, U) / \partial U \big|_{(Z,A,U)=(z_1,a,u_1)}.
\end{aligned} \tag{45}
$$

Then, we have

$$
\begin{aligned}
& f_{X_1, X_2, A | Z_1, Z_2}(x_1, x_2, a | z_1, z_2) \\
= & \, f_{A, U_1, U_2 | Z_1, Z_2}\left(a, g^{-1}(x_1, z_1, a), g^{-1}(x_2, z_2, a) \middle| z_1, z_2\right) |J_1| \\
= & \, f_{A, U_1, U_2 | Z_1, Z_2}\left(a, g^{-1}(x_2, z_2, a), g^{-1}(x_1, z_1, a) \middle| z_2, z_1\right) |J_1| \\
= & \, f_{X_1, X_2, A | Z_1, Z_2}(x_2, x_1, a | z_2, z_1) |J_2 J_1| \\
= & \, f_{X_1, X_2, A | Z_1, Z_2}(x_2, x_1, a | z_2, z_1),
\end{aligned} \tag{46}
$$

where the first equality holds by change of variables, the second equality uses (43) and $Z \perp (A, U)$, the latter of which enables one to switch the order of (z_1, z_2) in the

conditioned set, the third equality holds again by change of variables and

$$\begin{aligned} X_1 &= g\left(z_2, a, g^{-1}\left(x_2, z_2, a\right)\right) = x_2 \\ X_2 &= g\left(z_1, a, g^{-1}\left(x_1, z_1, a\right)\right) = x_1, \end{aligned} \tag{47}$$

and the last equality uses the fact that the product of derivatives of inverse functions is 1, i.e.,

$$\begin{aligned} & J_1 J_2 \\ &= \partial g^{-1}\left(X, Z, A\right) / \partial X\Big|_{(X,Z,A)=(x_1,z_1,a)} \times \partial g^{-1}\left(X, Z, A\right) / \partial X\Big|_{(X,Z,A)=(x_2,z_2,a)} \\ &\quad \times \partial g\left(Z, A, U\right) / \partial U\big|_{(Z,A,U)=(z_2,a,u_2)} \times \partial g\left(Z, A, U\right) / \partial U\big|_{(Z,A,U)=(z_1,a,u_1)} \\ &= \left[\partial g^{-1}\left(X, Z, A\right) / \partial X\Big|_{(X,Z,A)=(x_1,z_1,a)} \times \partial g\left(Z, A, U\right) / \partial U\big|_{(Z,A,U)=(z_1,a,u_1)}\right] \\ &\quad \times \left[\partial g^{-1}\left(X, Z, A\right) / \partial X\Big|_{(X,Z,A)=(x_2,z_2,a)} \times \partial g\left(Z, A, U\right) / \partial U\big|_{(Z,A,U)=(z_2,a,u_2)}\right] \\ &= 1 \times 1 = 1. \end{aligned} \tag{48}$$

Given (46), we have

$$\begin{aligned} f_{X_1,X_2|Z_1,Z_2}\left(x_1, x_2 \middle| z_1, z_2\right) &= \int f_{X_1,X_2,A|Z_1,Z_2}\left(x_1, x_2, a \middle| z_1, z_2\right) \mu\left(da\right) \\ &= \int f_{X_1,X_2,A|Z_1,Z_2}\left(x_2, x_1, a \middle| z_2, z_1\right) \mu\left(da\right) \\ &= f_{X_1,X_2|Z_1,Z_2}\left(x_2, x_1 \middle| z_2, z_1\right). \end{aligned} \tag{49}$$

which implies

$$\begin{aligned} & f_{A|X_1,X_2,Z_1,Z_2}\left(a \middle| x_1, x_2, z_1, z_2\right) \\ &= f_{X_1,X_2,A|Z_1,Z_2}\left(x_1, x_2, a \middle| z_1, z_2\right) / f_{X_1,X_2|Z_1,Z_2}\left(x_1, x_2 \middle| z_1, z_2\right) \\ &= f_{X_1,X_2,A|Z_1,Z_2}\left(x_2, x_1, a \middle| z_2, z_1\right) / f_{X_1,X_2|Z_1,Z_2}\left(x_2, x_1 \middle| z_2, z_1\right) \\ &= f_{A|X_1,X_2,Z_1,Z_2}\left(a \middle| x_2, x_1, z_2, z_1\right), \end{aligned} \tag{50}$$

where the second equality holds by (46) and (49).

Next, we follow Altonji and Matzkin (2005) to show that the conditional density $f_{A|X_1,X_2,Z_1,Z_2}\left(a \middle| x_1, x_2, z_1, z_2\right)$ can be approximated arbitrarily closely by a function of the form $f_{A|W}\left(a \middle| W\right)$, where W is a vector-valued function symmetric in the elements

of X and Z. By Assumption 3, the supports of X and Z are compact. By Assumption 1–3, $f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2)$ is continuous in (X_1, X_2, Z_1, Z_2). Therefore, from the Stone-Weierstrass Theorem one can find a function $f^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2)$ that is a polynomial in (X_1, X_2, Z_1, Z_2) over a compact set with the property that for any fixed δ that is arbitrarily close to 0,

$$\max_{x_t \in \mathcal{X}, z_t \in \mathcal{Z}, \forall t} \left| f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) - f^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) \right| \leq \delta. \quad (51)$$

Let

$$\begin{aligned} &\overline{f}_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) \\ &:= \left[f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) + f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_2, x_1, z_2, z_1) \right] / 2! \end{aligned} \quad (52)$$

denote the simple averages of $f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2)$ over all $T!$ (here $T = 2$) unique permutations of (x_t, z_t), and similarly for $\overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2)$. By (50), we have

$$\overline{f}_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) = f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2). \quad (53)$$

Also note that by construction, we have

$$\overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) = \overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_2, x_1, z_2, z_1). \quad (54)$$

By (50) and T, it is true that

$$\begin{aligned} &\left| f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) - \overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) \right| \\ = &\left| \overline{f}_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) - \overline{f}^w_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2) \right| \\ \leq &\, T! \times (\delta / T!) = \delta. \end{aligned} \quad (55)$$

Since f^w can be chosen to make δ arbitrarily small, (55) implies that $f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2)$ can be approximated arbitrarily closely by a polynomial $\overline{f}^w$ that is symmetric in (x_t, z_t) for $t = 1, 2$. Thus, by the fundamental theorem of symmetric functions, $\overline{f}^w$ can be written as a polynomial function of the elementary symmetric functions of $((x_1, z_1), (x_2, z_2))$. We denote this function by W and obtain that $f_{A|X_1,X_2,Z_1,Z_2}(a|\, x_1, x_2, z_1, z_2)$ can be approximated arbitrarily closely by

$f_{A|W}(a|W)$. Let $\delta \to 0$ in (51). Then, for any $t \in \{1,..,T\}$ and (X_t, Z_t, A, W) on its support we have

$$f_{A|X_t,Z_t,W}(a|x_t, z_t, w) = f_{A|W}(a|w). \tag{56}$$

To see why Assumption 1 only requires one coordinate of X_t to be strictly monotonic in U_t, suppose $X_t = (K_t, L_t)' = (g_K(Z_t, A, U_t), g_L(Z_t, A, U_t))'$ and only g_K is strictly monotonic in U_t. Then, to establish a similar result as (46), for $(k_1, l_1, k_2, l_2, z_1, z_2, a)$ on the support of $(K_1, L_1, K_2, L_2, Z_1, Z_2, A)$ we have

$$\begin{aligned}
& f_{K_1,L_1,K_2,L_2,A|Z_1,Z_2}(k_1, l_1, k_2, l_2, a|z_1, z_2) \\
= & f_{U_1,L_1,U_2,L_2,A|Z_1,Z_2}\left(g_K^{-1}(k_1, z_1, a), l_1, g_K^{-1}(k_2, z_2, a), l_2, a\middle| z_1, z_2\right)\left|\widetilde{J}_1\right| \\
= & f_{A,U_1,U_2|Z_1,Z_2}\left(a, g_K^{-1}(k_1, z_1, a), g_K^{-1}(k_2, z_2, a)\middle| z_1, z_2\right)\left|\widetilde{J}_1\right| \\
= & f_{A,U_1,U_2|Z_1,Z_2}\left(a, g_K^{-1}(k_2, z_2, a), g_K^{-1}(k_1, z_1, a)\middle| z_2, z_1\right)\left|\widetilde{J}_1\right| \\
= & f_{U_1,L_1,U_2,L_2,A|Z_1,Z_2}\left(g_K^{-1}(k_2, z_2, a), l_2, g_K^{-1}(k_1, z_1, a), l_1, a\middle| z_2, z_1\right)\left|\widetilde{J}_1\right| \\
= & f_{K_1,L_1,K_2,L_2,A|Z_1,Z_2}(k_2, l_2, k_1, l_1, a|z_2, z_1)\left|\widetilde{J}_2\right|\left|\widetilde{J}_1\right| \\
= & f_{K_1,L_1,K_2,L_2,A|Z_1,Z_2}(k_2, l_2, k_1, l_1, a|z_2, z_1),
\end{aligned} \tag{57}$$

where the first and second to last equality holds by change of variables, the second and fourth equality holds because L is a function of (Z, A, U), the third equality holds by (43) and the exogeneity of $Z \perp (A, U)$, and the last equality holds by $\left|\widetilde{J}_2\right|\left|\widetilde{J}_1\right| = 1$ which is derived similarly to (48). The rest of the proof follows similarly as in the scalar X case above. □

Proof of Lemma 2. Let $g^{-1}(x, z, a)$ denote the inverse function for $g(z, a, u)$ in its first argument, which exists by Assumption 1. Assume X_{it} is a scalar for brevity of exposition. For any (x, z, a, w) in the support of (X, Z, A, W), we have

$$\begin{aligned}
& F_{X_{it}|Z_{it},W_i}(x|z, w) \\
= & F_{X_{it}|Z_{it},A_i,W_i}(x|z, a, w) \\
= & \mathbb{P}(X_{it} \le x|Z_{it} = z, A_i = a, W_i = w) \\
= & \mathbb{P}(g(z, a, U_{it}) \le x|Z_{it} = z, A_i = a, W_i = w) \\
= & \mathbb{P}\left(U_{it} \le g^{-1}(x, z, a)\middle| A_i = a, W_i = w\right)
\end{aligned}$$

$$= F_{U_{it}|A_i,W_i}\left(g^{-1}\left(x,z,a\right)\middle|\, a,w\right), \tag{58}$$

where the first equality holds by (12), the third uses (3), the fourth holds by Assumption 1 and 4, and the last equality holds by definition of the conditional CDF of U_{it} given (A_i, W_i).

By (3), $U_{it} = g^{-1}\left(X_{it}, Z_{it}, A_i\right)$, so that plugging in gives

$$V_{it} \coloneqq F_{X_{it}|Z_{it},W_i}\left(X_{it}|\, Z_{it}, W_i\right) = F_{U_{it}|A_i,W_i}\left(U_{it}|\, A_i, W_i\right). \tag{59}$$

□

B Proofs in Section 4

The proof of Lemma 3 follows directly from that of Theorem 12 in Imbens and Newey (2009). Thus, it is omitted for brevity. First, we prove Theorem 2. Note that by T, we obtain the mean squared and uniform convergence results if we can prove it for each coordinate of β. Therefore, wlog we assume β is a scalar throughout the proof. Then, we prove Theorem 3 and 4. The proof of Theorem 3 follows from Imbens and Newey (2002), Andrews (1991), and a Cramér–Wold device. The proof of Theorem 4 requires more efforts. As discussed before, for $\overline{\beta}$ one can obtain its normality by choosing the basis function $r^M(\cdot) \equiv 1$ and applying the results for $\beta(x)$.

Proof of Theorem 2. As discussed before, the convergence rate for $\widehat{\beta}(v,w)$ is the same as $\widehat{G}(s)$ because they share the same series regression coefficients $\widehat{\alpha}^K$. Under Assumption 7 and 8, the convergence rate result on $\widehat{G}(s)$ applies directly to $\widehat{\beta}(v,w)$ and the proof is thus omitted.

We focus on $\widehat{\beta}(x)$, since the result for $\widehat{\overline{\beta}}$ follows by setting $r^M(\cdot) \equiv 1$. Following Newey (1997), we normalize $\mathbb{E} r_i r_i' = I$ and have $\lambda_{\min}\left(\widehat{R}\right) \geq C > 0$. By (35), we have

$$\begin{aligned}
&\left\|\widehat{R}^{1/2}\left(\widehat{\eta}^M - \eta^M\right)\right\|^2 \\
\leq &\left(\widehat{B} - \widetilde{B}\right)' r\widehat{R}^{-1}r'\left(\widehat{B} - \widetilde{B}\right)/n^2 + \left(\widetilde{B} - B\right)' r\widehat{R}^{-1}r'\left(\widetilde{B} - B\right)/n^2 \\
&+ \left(B - B^X\right)' r\widehat{R}^{-1}r'\left(B - B^X\right)/n^2 + \left(B^X - r\eta^M\right)' r\widehat{R}^{-1}r'\left(B^X - r\eta^M\right)/n^2.
\end{aligned} \tag{60}$$

Following the proof for Theorem 1 of Newey (1997), Lemma A1 and Lemma A3 of Imbens and Newey (2002), under Assumption 7 we have

$$\left\| n^{-1} \sum_i \widehat{\overline{p}}_i \widehat{\overline{p}}_i^{'} - \mathbb{E}\overline{p}_i \overline{p}_i^{'} \right\| = o_P(1) \text{ and } \mathbb{E}\overline{p}_i \overline{p}_i^{'} \leq CI. \tag{61}$$

Then, we have

$$\begin{aligned}
&\left(\widehat{B} - \widetilde{B}\right)^{'} r \widehat{R}^{-1} r^{'} \left(\widehat{B} - \widetilde{B}\right) / n^2 \\
&\leq C \left(\widehat{B} - \widetilde{B}\right)^{'} \left(\widehat{B} - \widetilde{B}\right) / n \\
&= C n^{-1} \sum_i \left(\widehat{\beta}\left(\widehat{V}_i, W_i\right) - \beta\left(\widehat{V}_i, W_i\right)\right)^2 \\
&= C n^{-1} \sum_i \left(\widehat{\overline{p}}_i^{'} \left(\widehat{\alpha}^K - \alpha^K\right) + \left(\widehat{\overline{p}}_i^{'} \alpha^K - \beta\left(\widehat{V}_i, W_i\right)\right)\right)^2 \\
&\leq C \left\|\widehat{\alpha}^K - \alpha^K\right\|^2 + \sup_{s \in \mathcal{S}} \left\|\overline{p}^K(s)^{'} \alpha^K - \beta(v, w)\right\|^2 = O_P\left(\Delta_{2n}^2\right)
\end{aligned} \tag{62}$$

where the first inequality holds because $r\widehat{R}^{-1}r^{'}/n$ is idempotent, the last inequality holds by (61), and the last equality uses Lemma 3.

Next, we have

$$\begin{aligned}
&\left(\widetilde{B} - B\right)^{'} r \widehat{R}^{-1} r^{'} \left(\widetilde{B} - B\right) / n^2 \\
&\leq C n^{-1} \sum_i \left(\beta\left(\widehat{V}_i, W_i\right) - \beta\left(V_i, W_i\right)\right)^2 \\
&\leq C n^{-1} \sum_i \left(\widehat{V}_i - V_i\right)^2 = O_P\left(\Delta_{1n}^2\right),
\end{aligned} \tag{63}$$

where the last inequality holds by Assumption 8(4) and the equality holds by Lemma 3.

Finally, for the last two terms in (60), we have

$$\begin{aligned}
&\mathbb{E}\left[\left(B - B^X\right)^{'} r \widehat{R}^{-1} r^{'} \left(B - B^X\right) / n^2 \middle| \mathbf{X}\right] \\
&= tr\left\{\mathbb{E}\left[\xi^{'} r \widehat{R}^{-1} r^{'} \xi \middle| \mathbf{X}\right]\right\} / n^2 \\
&= tr\left\{\mathbb{E}\left[\xi\xi^{'} \middle| \mathbf{X}\right] r \widehat{R}^{-1} r^{'}\right\} / n^2 \\
&\leq tr\left\{CIr\widehat{R}^{-1}r^{'}\right\} / n^2 = Ctr\left\{\widehat{R}^{-1}\widehat{R}\right\} / n = CM/n.
\end{aligned} \tag{64}$$

and

$$\begin{aligned}&\left(B^X - R\eta^M\right)' r\widehat{R}^{-1}r'\left(B^X - R\eta^M\right)/n^2\\ \leq &\left(B^X - R\eta^M\right)'\left(B^X - R\eta^M\right)/n = O_P\left(M^{-2d_3}\right).\end{aligned} \tag{65}$$

Collecting terms and using $\lambda_{\min}\left(\widehat{R}\right) \geq C$, we have

$$\left\|\widehat{\eta}^M - \eta^M\right\|^2 = O_P\left(\Delta_{2n}^2 + M/n + M^{-2d_3}\right) =: O_P\left(\Delta_{3n}^2\right), \tag{66}$$

which implies

$$\begin{aligned}&\int \left\|\widehat{\beta}(x) - \beta(x)\right\|^2 dF(x)\\ \leq &\int \left(r^M(x)'\left(\widehat{\eta}^M - \eta^M\right) + \left(r^M(x)'\eta^M - \beta(x)\right)\right)^2 dF(x)\\ \leq &\, C\left\|\widehat{\eta}^M - \eta^M\right\|^2 + \sup_{x\in\mathcal{X}}\left|\beta(x) - r^M(x)'\eta^M\right|^2 = O_P\left(\Delta_{3n}^2\right),\end{aligned} \tag{67}$$

and

$$\begin{aligned}\sup_{x\in\mathcal{X}}\left\|\widehat{\beta}(x) - \beta(x)\right\| &\leq \sup_{x\in\mathcal{X}}\left\|r^M(x)\right\|\left\|\widehat{\eta}^M - \eta^M\right\| + \sup_{x\in\mathcal{X}}\left|\beta(x) - r^M(x)'\eta^M\right|\\ &= O_P\left(\zeta(M)\Delta_{3n}\right).\end{aligned}$$

□

Proof of Theorem 3. Recall that the analysis of Imbens and Newey (2002) applies to scalar functionals of $G(s)$. By Cramér–Wold device and Imbens and Newey (2002), for any constant vector c with $c'c = 1$ we have

$$\begin{aligned}&c'\sqrt{n}\Omega_1^{-1/2}\left(\widehat{\beta}(v,w) - \beta(v,w)\right) \to_d N(0,1) \text{ and}\\ &\left(c'\Omega_1 c\right)^{-1}\left[c'\left(\widehat{\Omega}_1 - \Omega_1\right)c\right] \xrightarrow{p} 0.\end{aligned} \tag{68}$$

By (68) and Assumption 9(6), it is true that

$$c'\left(b_{1n}\widehat{\Omega}_1 - b_{1n}\Omega_1\right)c \xrightarrow{p} 0, \tag{69}$$

which implies

$$b_{1n}\widehat{\Omega}_1 \xrightarrow{p} \overline{\Omega}_1. \tag{70}$$

Combining (68) – (70), we have

$$\begin{aligned}
&\sqrt{n}\widehat{\Omega}_1^{-1/2}\left(\widehat{\beta}(v,w)-\beta(v,w)\right)\\
=&\left(b_{1n}\widehat{\Omega}_1\right)^{-1/2}(b_{1n}\Omega_1)^{1/2}\sqrt{n}\Omega_1^{-1/2}\left(\widehat{\beta}(v,w)-\beta(v,w)\right)\\
\xrightarrow{d}&\overline{\Omega}_1^{-1/2}\overline{\Omega}_1^{1/2}\mathcal{N}(0,I)=\mathcal{N}(0,I),
\end{aligned} \tag{71}$$

where the convergence holds by (68), (70), and Assumption 9(6). □

Proof of Theorem 4. Following the proof of Theorem 3, one can extend the results to vector-valued functionals using Cramér–Wold device and the proofs of Andrews (1991). Therefore, wlog we assume $\beta(x)$ is a scalar in this proof. First, we derive the influence functions that correctly account for the effects from estimating $\beta(x)$ and prove asymptotic normality using Lindeberg–Feller CLT. Then, we show consistency for the estimator of the variance, which can be used to construct feasible confidence intervals. We write $r^M(x)$ as $r(x)$ and suppress t subscript when there is no confusion.

By Assumption 10(1), we normalize $\mathbb{E}r_ir_i' = I$ and obtain $\left\|\widehat{R}-I\right\| = o_P(1)$ using a similar argument as in the proof of Theorem 1 of Newey (1997). Recall that $\widehat{\beta}(x) = r^M(x)'\widehat{R}^{-1}r'\widehat{B}/n$. Let

$$\widehat{a}\left(\widehat{\beta},\widehat{V}\right) = r^M(x)'\widehat{R}^{-1}r'\widehat{B}/n, \text{ and } a(\beta,V) = \mathbb{E}\left[\beta_i|\,X=x\right] \tag{72}$$

and define

$$\begin{aligned}
\Omega_{21} &= \mathbb{E}\left(A_1P^{-1}p_iu_i\right)\left(A_1P^{-1}p_iu_i\right)'\\
\Omega_{22} &= \mathbb{E}\left[\begin{array}{c}\left(A_1P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II}+r_i\left(\beta(V_i,W_i)-\beta(X_i)\right)\right)\right)\\ \times\left(A_1P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II}+r_i\left(\beta(V_i,W_i)-\beta(X_i)\right)\right)\right)'\end{array}\right].
\end{aligned} \tag{73}$$

Then, we have $\Omega_2 = \Omega_{21}+\Omega_{22}$.

Let $F = \Omega_2^{-1/2}$, which is well-defined because

$$\begin{aligned}
\Omega_{21} &= A_1P^{-1}\left(\mathbb{E}p_ip_i'u_i^2\right)P^{-1}A_1'\\
&= A_1P^{-1}\left(\mathbb{E}p_ip_i'\mathbb{E}\left(u_i^2\middle|\,X_i,V_i,W_i\right)\right)P^{-1}A_1'
\end{aligned}$$

$$\geq CA_1P^{-1}A_1' = Cr(x)' \left(\mathbb{E}r_i\overline{p}_i'\right)\left(\mathbb{E}p_ip_i'\right)^{-1}\left(\mathbb{E}\overline{p}_ir_i'\right)r(x) > 0, \tag{74}$$

where the first inequality holds by Assumption 9(3) and the last inequality holds by Assumption 10(1).

We expand

$$\begin{aligned}
&\sqrt{n}F\left(\widehat{a}\left(\widehat{\beta},\widehat{V}\right) - a(\beta,V)\right)\\
&= \sqrt{n}F\left(\widehat{a}\left(\widehat{\beta},\widehat{V}\right) - \widehat{a}\left(\beta,\widehat{V}\right) + \widehat{a}\left(\beta,\widehat{V}\right) - \widehat{a}(\beta,V) + \widehat{a}(\beta,V) - a(\beta,V)\right)\\
&= n^{-1/2}\sum_i(\psi_{1i}+\psi_{2i}+\psi_{3i}) + o_P(1)
\end{aligned} \tag{75}$$

and show that

$$\psi_{1i} = H_1\left(p_iu_i - \overline{\mu}_i^I\right),\ \psi_{2i} = H_2\overline{\mu}_i^{II},\ \text{and } \psi_{3i} = H_2r_i\xi_i. \tag{76}$$

First, for ψ_{1i} we have

$$\begin{aligned}
&\sqrt{n}F\left(\widehat{a}\left(\widehat{\beta},\widehat{V}\right) - \widehat{a}\left(\beta,\widehat{V}\right)\right)\\
&= \sqrt{n}Fr(x)'\widehat{R}^{-1}r'\left(\widehat{B}-\widetilde{B}\right)/n\\
&= \sqrt{n}Fr(x)'\widehat{R}^{-1}r'\left(\widehat{\overline{p}}\widehat{P}^{-1}\widehat{p}'Y/n - \widetilde{B}\right)/n\\
&= n^{-1/2}Fr(x)'\widehat{R}^{-1}r'\left[n^{-1}\widehat{\overline{p}}\widehat{P}^{-1}\widehat{p}'\left(Y-G+G-\widetilde{G}+\widetilde{G}-\widehat{p}\alpha^K\right) + \left(\widehat{\overline{p}}\alpha^K - \widetilde{B}\right)\right]\\
&= n^{-1/2}\sum_i\widehat{H}_1\widehat{p}_i\left[u_i - \left(G(\widehat{s}_i) - G(s_i)\right)\right] + n^{-1/2}\widehat{H}_1\widehat{p}'\left(\widetilde{G}-\widehat{p}\alpha^K\right)\\
&\quad + n^{-1/2}\widehat{H}_2r'\left(\widehat{\overline{p}}\alpha^K - \widetilde{B}\right) =: D_{11}+D_{12}+D_{13}.
\end{aligned} \tag{77}$$

We show $D_{11} = n^{-1/2}\sum_i\psi_{1i} + o_P(1)$, $D_{12} = o_P(1)$, and $D_{13} = o_P(1)$.

The proof of

$$D_{11} = n^{-1/2}\sum_i\psi_{1i} + o_P(1) \tag{78}$$

is analogous to that of Lemma B7 and B8 of Imbens and Newey (2002), except that we need to establish $\left\|\widehat{H}_1 - H_1\right\| = o_P(1)$. To prove this claim, first we have

$$\|H_1\| = O(1)\ \text{ and }\ \|H_2\| = O(1), \tag{79}$$

because $\|H_1\|^2 \leq CA_1A_1'/\Omega_2 \leq C$ and $\|H_2\|^2 = A_2A_2'/\Omega_2 \leq CA_1A_1'/\Omega_2 \leq C$. In addi-

tion, we have $\left\|\widehat{P}-P\right\|=o_P(1)$, $\left\|\widehat{R}-I\right\|=o_P(1)$, and $\left\|n^{-1}\sum_i r_i\overline{p}_i-\mathbb{E}r_i\overline{p}_i'\right\|=o_P(1)$ as in the proof of Theorem 1 of Newey (1997). By Slutsky Theorem, $\left\|\widehat{R}^{-1}-I\right\|=o_P(1)$. Using CS and Lemma A3 of Imbens and Newey (2002), we have

$$\begin{aligned}\left\|n^{-1}\sum_i r_i\left(\overline{p}_i-\widehat{\overline{p}}_i\right)'\right\|^2 &\leq n^{-1}\sum_i\|r_i\|^2\times n^{-1}\sum_i\left\|\widehat{\overline{p}}_i-\overline{p}_i\right\|^2\\ &=O_P\left(M\zeta_1(K)^2\Delta_n^2\right)=o_P(1).\end{aligned}\tag{80}$$

Therefore, by T we have with probability approaching 1

$$\begin{aligned}&\left\|\widehat{H}_1-H_1\right\|^2\\ =&\left\|F\widehat{A}_1\widehat{P}^{-1}-FA_1P^{-1}\right\|^2\\ \leq& 2\left\|F\left(\widehat{A}_1-A_1\right)\widehat{P}^{-1}\right\|^2+2\left\|FA_1\left(\widehat{P}^{-1}-P^{-1}\right)\right\|^2\\ =& 2\left\|F\left(r(x)'(I+o_P(1))\left(\mathbb{E}r_i\overline{p}_i'+o_P(1)\right)-r(x)'\mathbb{E}r_i\overline{p}_i'\right)\widehat{P}^{-1}\right\|^2\\ &+2\left\|FA_1P^{-1}\left(P-\widehat{P}\right)\widehat{P}^{-1}\right\|^2\\ \leq&\|H_2\|^2o_P(1)+\|H_1\|^2o_P(1)=o_P(1).\end{aligned}\tag{81}$$

and similarly $\left\|\widehat{H}_2-H_2\right\|=o_P(1)$. The result follows as in the proof of Lemma B7 and B8 of Imbens and Newey (2002).

Next, recall that

$$\left(\widetilde{G}-\widehat{p}\alpha^K\right)'\left(\widetilde{G}-\widehat{p}\alpha^K\right)/n=O_P\left(K^{-2d_2}\right)\tag{82}$$

by Assumption 7(4). Therefore,

$$\begin{aligned}\left|n^{-1/2}\widehat{H}_1\widehat{p}'\left(\widetilde{G}-\widehat{p}\alpha^K\right)\right|^2&\leq n\left[\widehat{H}_1\widehat{P}\widehat{H}_1'\right]\left[\left(\widetilde{G}-\widehat{p}\alpha^K\right)'\left(\widetilde{G}-\widehat{p}\alpha^K\right)/n\right]\\ &\leq\left\|\widehat{H}_1\right\|^2O_P\left(nK^{-2d_2}\right)=o_P(1).\end{aligned}\tag{83}$$

For D_{13}, similarly to (83) we have

$$\begin{aligned}\left|n^{-1/2}\widehat{H}_2r'\left(\widehat{p}\alpha^K-\widetilde{B}\right)\right|^2&\leq n\left[\widehat{H}_2\widehat{R}\widehat{H}_2\right]\left[\left(\widetilde{B}-\widehat{p}\alpha^K\right)'\left(\widetilde{B}-\widehat{p}\alpha^K\right)/n\right]\\ &=O_P\left(nK^{-2d}\right)=o_P(1).\end{aligned}\tag{84}$$

Summarizing (78)–(84), we obtain

$$\psi_{1i} = H_1 \left(p_i u_i - \overline{\mu}_i^I \right). \tag{85}$$

To obtain ψ_{2i}, we have

$$\begin{aligned}
&\sqrt{n} F \left(\widehat{a} \left(\beta, \widehat{V} \right) - \widehat{a} \left(\beta, V \right) \right) \\
&= \sqrt{n} F r \left(x \right)' \widehat{R}^{-1} r' \left(\widetilde{B} - B \right) / n \\
&= \widehat{H}_2 n^{-1/2} \sum_i r_i \left(\widetilde{\beta}_i - \beta_i \right) \\
&= \widehat{H}_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\widehat{V}_i - V_i \right) + \widehat{H}_2 n^{-1/2} \sum_i r_i \beta_{vv} \left(\widetilde{V}_i, W_i \right) \left(\widehat{V}_i - V_i \right)^2 / 2 \\
&=: D_{21} + D_{22}.
\end{aligned} \tag{86}$$

We prove $D_{21} = n^{-1/2} \sum_i H_2 \overline{\mu}_i^{II} + o_P(1)$ and $D_{22} = o_P(1)$. For D_{21}, we have

$$\begin{aligned}
D_{21} &= \widehat{H}_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\widehat{V}_i - V_i \right) \\
&= H_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \Delta_i^I + \left(\widehat{H}_2 - H_2 \right) n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\widehat{V}_i - V_i \right) \\
&\quad + H_2 n^{-1/2} \sum_i r_i \beta_v \left(V_i, W_i \right) \left(\Delta_i^{II} + \Delta_i^{III} \right) \\
&=: D_{211} + D_{212} + D_{213},
\end{aligned} \tag{87}$$

where

$$\begin{aligned}
&\delta_{ij} = F \left(X_i | \, Z_j, W_j \right) - q_j' \gamma^L \left(X_i \right), \ \Delta_i^I = q_i' \widehat{Q}^{-1} \sum_j q_j v_{ij} / n, \\
&\Delta_i^{II} = q_i' \widehat{Q}^{-1} \sum_j q_j \delta_{ij} / n, \text{ and } \Delta_i^{III} = -\delta_{ii}.
\end{aligned} \tag{88}$$

Following the proof of Lemma B7 of Imbens and Newey (2002), we obtain

$$D_{211} = n^{-1/2} \sum_i H_2 \overline{\mu}_i^{II} + o_P(1). \tag{89}$$

For D_{212}, we have

$$\begin{aligned}|D_{212}|^2 &\leq Cn\left[\left(\widehat{H}_2 - H_2\right)\widehat{R}\left(\widehat{H}_2 - H_2\right)'\right]\left[n^{-1}\sum_i\left(\widehat{V}_i - V_i\right)^2\right] \\ &= O_P\left\{n\left(\zeta\left(M\right)^2 M/n\right)\Delta_{1n}^2\right\} = o_P\left(1\right). \end{aligned} \tag{90}$$

For D_{213}, we have

$$|D_{213}|^2 \leq Cn\left[H_2\widehat{R}H_2'\right]\left[\sum_i\left(\left(\Delta_i^{II}\right)^2 + \left(\Delta_i^{III}\right)^2\right)/n\right] = O_P\left(nL^{1-2d_1}\right) = o_P\left(1\right), \tag{91}$$

where the first equality is established in the proof of Theorem 4 of Imbens and Newey (2002).

Next, for D_{22}, we have

$$\begin{aligned}|D_{22}| &\leq C\sqrt{n}\left\|\widehat{H}_2\right\|\sup_{x\in\mathcal{X}}\left\|r\left(x\right)\right\|\left|n^{-1}\sum_i\left(\widehat{V}_i - V_i\right)^2\right| \\ &= O_P\left(\sqrt{n}\zeta\left(M\right)\Delta_n^2\right) = o_P\left(1\right). \end{aligned} \tag{92}$$

Combining the results for D_{21} and D_{22}, we obtain

$$\sqrt{n}F\left(\widehat{a}\left(\beta,\widehat{V}\right) - \widehat{a}\left(\beta,V\right)\right) = n^{-1/2}\sum_i H_2\overline{\mu}_i^{II} + o_P\left(1\right). \tag{93}$$

To obtain ψ_{3i}, first we expand

$$\begin{aligned}&\sqrt{n}F\left(\widehat{a}\left(\beta,V\right) - a\left(\beta,V\right)\right) \\ &= n^{-1/2}\sum_i\widehat{H}_2 r_i\beta_i - \sqrt{n}F\beta\left(x\right) \\ &= n^{-1/2}\sum_i H_2 r_i\left(\beta\left(V_i,W_i\right) - \beta\left(X_i\right)\right) + n^{-1/2}\sum_i\left(\widehat{H}_2 - H_2\right)r_i\left(\beta\left(V_i,W_i\right) - \beta\left(X_i\right)\right) \\ &\quad + n^{-1/2}\sum_i\widehat{H}_2 r_i\left(\beta\left(X_i\right) - r_i'\eta^M\right) - \sqrt{n}F\left(\beta\left(x\right) - r\left(x\right)'\eta^M\right) \\ &=: D_{31} + D_{32} + D_{33} + D_{34}. \end{aligned} \tag{94}$$

Recall that $D_{31} = n^{-1/2}\sum_i H_2 r_i\xi_i$ by definition of ξ_i. Thus, we show D_{32}, D_{33}, and D_{34} are all $o_P\left(1\right)$.

For D_{32}, we have

$$\begin{aligned}\mathbb{E}\left[\left|D_{32}\right|^2\middle|\mathbf{X}\right] &= \left(\widehat{H}_2 - H_2\right) r' \mathbb{E}\left[\xi\xi'\middle|\mathbf{X}\right] r \left(\widehat{H}_2 - H_2\right)' /n \\ &\leq C\left(\widehat{H}_2 - H_2\right)\widehat{R}\left(\widehat{H}_2 - H_2\right)' \\ &\leq C\left\|\widehat{H}_2 - H_2\right\|^2\left(1 + \left\|\widehat{R} - I\right\|\right) \\ &= O_P\left\{\left\|\widehat{H}_2 - H_2\right\|^2\right\} = O_P\left(\zeta\left(M\right)^2 M/n\right) = o_P\left(1\right), \end{aligned} \tag{95}$$

where the first inequality holds by Assumption 8(3) and the fact that $\widehat{H}_2$ and r are functions of X_i only, the second equality holds by $\left\|\widehat{R} - I\right\| = o_P\left(1\right)$, and the third equality follows similarly as in equation (A.1) and (A.6) of Newey (1997). Therefore, $D_{32} = o_P\left(1\right)$ by CM.

For D_{33}, by CS we have

$$\begin{aligned}\left|D_{33}\right|^2 &\leq n\left(\widehat{H}_2\widehat{R}\widehat{H}_2'\right)\sum_i\left(\beta\left(X_i\right) - r_i'\eta^M\right)^2/n \\ &= O_P\left(nM^{-2d_3}\right) = o_P\left(1\right), \end{aligned} \tag{96}$$

where the first equality holds by Assumption 8(1).

For D_{34}, we have

$$\left|D_{34}\right|^2 = nF^2\left(\beta\left(x\right) - r\left(x\right)'\eta^M\right)^2 = O_P\left(nM^{-2d_3}\right) = o_P\left(1\right). \tag{97}$$

Summarizing (94)–(97), we obtain

$$\sqrt{n}F\left(\widehat{a}\left(\beta, V\right) - a\left(\beta, V\right)\right) = n^{-1/2}\sum_i H_2 r_i \xi_i + o_P\left(1\right). \tag{98}$$

In sum, we have shown

$$\sqrt{n}F\left(\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - a\left(\beta, V\right)\right) = n^{-1/2}\sum_i\left(\psi_{1i} + \psi_{2i} + \psi_{3i}\right) + o_P\left(1\right), \tag{99}$$

where

$$\psi_{1i} = H_1\left(p_i u_i - \overline{\mu}_i^I\right),\ \psi_{2i} = H_2\overline{\mu}_i^{II},\ \text{and}\ \psi_{3i} = H_2 r_i \xi_i \tag{100}$$

and

$$H_1 p_i u_i \perp \left(H_1 \overline{\mu}_i^I, H_2 \overline{\mu}_i^{II}, H_2 r_i \xi_i\right) \tag{101}$$

because $\mathbb{E}\left(u_i \middle| X_i, V_i, W_i\right) = 0$ by construction.

Let $\Psi_{in} = n^{-1/2}\left(\psi_{1i} + \psi_{2i} + \psi_{3i}\right)$. We have $\mathbb{E}\Psi_{in} = 0$ and $Var\left(\Psi_{in}\right) = 1/n$. For any $\varepsilon > 0$, under Assumption 9 and 10, we have

$$\begin{aligned}
& n\mathbb{E}\left[\mathbb{1}\left\{\left|\Psi_{in}\right| > \varepsilon\right\}\Psi_{in}^2\right] \\
\leq\ & n\varepsilon^2\mathbb{E}\left[\mathbb{1}\left\{\left|\Psi_{in}\right| > \varepsilon\right\}\left(\Psi_{in}/\varepsilon\right)^4\right] \leq n\varepsilon^{-2}\mathbb{E}\Psi_{in}^4 \\
\leq\ & C\mathbb{E}\left[\left(H_1 p_i u_i\right)^4 + \left(H_1\overline{\mu}_i^I\right)^4 + \left(H_2\overline{\mu}_i^{II}\right)^4 + \left(H_2 r_i \xi_i\right)^4\right]/n \\
\leq\ & C\left(\zeta\left(K\right)^2 K + \zeta\left(K\right)^4\zeta\left(L\right)^4 L + \zeta\left(M\right)^4\zeta\left(L\right)^4 L + \zeta\left(M\right)^2 M\right)/n \to 0,
\end{aligned} \tag{102}$$

where the last inequality follows a similar argument as in the proof of Lemma B5 of Imbens and Newey (2002). Then, by Lindeberg–Feller CLT we obtain

$$\sqrt{n}\Omega_2^{-1/2}\left(\widehat{a}\left(\widehat{\beta}, \widehat{V}\right) - a\left(\beta, V\right)\right) \xrightarrow{d} N\left(0, 1\right). \tag{103}$$

To construct a feasible confidence interval, one needs a consistent estimator of the covariance matrix. Thus, we show $\widehat{\Omega}_2/\Omega_2 - 1 \xrightarrow{p} 0$. Recall that

$$\Omega_2 = \mathbb{E}\left(A_1 P^{-1} p_i u_i\right)^2 + \mathbb{E}\left(A_1 P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i\xi_i\right)\right)^2 = \Omega_{21} + \Omega_{22} \tag{104}$$

and

$$\widehat{\Omega}_2 = n^{-1}\sum_i\left(\widehat{A}_1\widehat{P}^{-1}\widehat{p}_i\widehat{u}_i\right)^2 + n^{-1}\sum_i\left(\widehat{A}_1\widehat{P}^{-1}\widehat{\mu}_i^I - \widehat{A}_2\widehat{R}^{-1}\left(\widehat{\mu}_i^{II} + r_i\widehat{\xi}_i\right)\right)^2 =: \widehat{\Omega}_{21} + \widehat{\Omega}_{22}. \tag{105}$$

The proof of $\widehat{\Omega}_{21}/\Omega_2 - \Omega_{21}/\Omega_2 \xrightarrow{p} 0$ follows the proof of Lemma B10 of Imbens and Newey (2009), with the $\widehat{A}_1$ instead of A_1 appearing in the definition of $\widehat{H}_1$. Nonetheless, we have shown that $\left\|\widehat{H}_1 - H_1\right\| = o_P\left(1\right)$. Thus, the proof for $\widehat{\Omega}_{21}$ follows similarly and is omitted for brevity.

For $\widehat{\Omega}_{22}$, we first show

$$n^{-1}\sum_i\left(\widehat{H}_1\widehat{\overline{\mu}}_i^I - H_1\overline{\mu}_i^I\right)^2 = o_P\left(1\right)$$

$$n^{-1}\sum_i\left(\widehat{H}_2\widehat{\overline{\mu}}_i^{II}-H_2\overline{\mu}_i^{II}\right)^2=o_P\left(1\right)$$
$$n^{-1}\sum_i\left(\widehat{H}_2r_i\widehat{\xi}_i-H_2r_i\xi_i\right)^2=o_P\left(1\right). \tag{106}$$

The first two convergence results hold by following the argument of the proof of Lemma B9 in Imbens and Newey (2002). For the last one, we have

$$\begin{aligned}
&\widehat{H}_2r_i\widehat{\xi}_i-H_2r_i\xi_i\\
&=\widehat{H}_2r_i\left(\widehat{\xi}_i-\xi_i\right)+\left(\widehat{H}_2-H_2\right)r_i\xi_i\\
&=\widehat{H}_2r_i\left(\widehat{\beta}\left(\widehat{V}_i,W_i\right)-\widehat{\beta}\left(X_i\right)-\beta\left(V_i,W_i\right)+\beta\left(X_i\right)\right)+\left(\widehat{H}_2-H_2\right)r_i\xi_i\\
&=\widehat{H}_2r_i\left(\widehat{\beta}\left(\widehat{V}_i,W_i\right)-\beta\left(\widehat{V}_i,W_i\right)\right)+\widehat{H}_2r_i\left(\beta\left(\widehat{V}_i,W_i\right)-\beta\left(V_i,W_i\right)\right)\\
&\quad+\widehat{H}_2r_i\left(\beta\left(X_i\right)-\widehat{\beta}\left(X_i\right)\right)+\left(\widehat{H}_2-H_2\right)r_i\xi_i\\
&=:D_{41i}+D_{42i}+D_{43i}+D_{44i}.
\end{aligned} \tag{107}$$

For D_{41}, we have

$$\begin{aligned}
n^{-1}\sum_i D_{41i}^2&\leq\left\|\widehat{H}_2\right\|^2\sup_{x\in\mathcal{X}}\left\|r\left(x\right)\right\|^2n^{-1}\sum_i\left(\widehat{\beta}\left(\widehat{V}_i,W_i\right)-\beta\left(\widehat{V}_i,W_i\right)\right)^2\\
&\leq C\zeta\left(M\right)^2n^{-1}\sum_i\left[\left(\widehat{\overline{p}}_i'\left(\widehat{\alpha}^K-\alpha^K\right)\right)^2+\left(\widehat{\overline{p}}_i'\alpha^K-\beta\left(\widehat{v}_i,w_i\right)\right)^2\right]\\
&=O_P\left(\zeta\left(M\right)^2\Delta_{2n}^2\right)=o_P\left(1\right),
\end{aligned} \tag{108}$$

where the second inequality holds by $\left\|\widehat{H}_2\right\|=O_P\left(1\right)$ and Assumption 10(1) and the first equality holds by (62).

For D_{42}, we have

$$\begin{aligned}
n^{-1}\sum_i D_{42i}^2&\leq\left\|\widehat{H}_2\right\|^2\sup_{x\in\mathcal{X}}\left\|r\left(x\right)\right\|^2n^{-1}\sum_i\left(\beta\left(\widehat{V}_i,W_i\right)-\beta\left(V_i,W_i\right)\right)^2\\
&\leq C\zeta\left(M\right)^2n^{-1}\sum_i\left(\widehat{V}_i-V_i\right)^2=O_P\left(\zeta\left(M\right)^2\Delta_{1n}^2\right)=o_P\left(1\right),
\end{aligned} \tag{109}$$

where the first equality holds by Lemma 3.

The proof of $n^{-1}\sum_i D_{43i}^2=o_P\left(1\right)$ is completely analogous to (108) and is thus omitted.

For D_{44}, we have

$$\begin{aligned}\mathbb{E}\left[n^{-1}\sum_i D_{44i}^2\middle|\mathbf{X}\right] &= \left(\widehat{H}_2 - H_2\right) n^{-1}\sum_i r_i r_i' \mathbb{E}\left(\xi_i^2\middle| X_i\right)\left(\widehat{H}_2 - H_2\right)' \\ &\leq C\left(\widehat{H}_2 - H_2\right)\widehat{R}\left(\widehat{H}_2 - H_2\right)' \\ &\leq C\left\|\widehat{H}_2 - H_2\right\|^2 = o_P(1), \end{aligned} \tag{110}$$

where the first equality holds by $\widehat{H}_2$ and r_i are both functions of $\mathbf{X}$, the first inequality holds by Assumption 8(3), and the last inequality uses $\left\|\widehat{R} - I\right\| = o_P(1)$. Then, by CM, we have

$$n^{-1}\sum_i D_{44i}^2 = o_P(1). \tag{111}$$

Combining results for D_{41}–D_{44}, we have

$$n^{-1}\sum_i \left(\widehat{H}_2 r_i \widehat{\xi}_i - H_2 r_i \xi_i\right)^2 = o_P(1). \tag{112}$$

Therefore, we have proven (106), which implies

$$\begin{aligned} & n^{-1}\sum_i \left(\left(\widehat{H}_1\widehat{\overline{\mu}}_i^I - \widehat{H}_2\widehat{\overline{\mu}}_i^{II} - \widehat{H}_2 r_i\widehat{\xi}_i\right) - \left(H_1\overline{\mu}_i^I - H_2\overline{\mu}_i^{II} - H_2 r_i\xi_i\right)\right)^2 \\ \leq\ & Cn^{-1}\sum_i \left(\widehat{H}_1\widehat{\overline{\mu}}_i^I - H_1\overline{\mu}_i^I\right)^2 + Cn^{-1}\sum_i \left(\widehat{H}_2\widehat{\overline{\mu}}_i^{II} - H_2\overline{\mu}_i^{II}\right)^2 \\ & + Cn^{-1}\sum_i \left(\widehat{H}_2 r_i\widehat{\xi}_i - H_2 r_i\xi_i\right)^2 = o_P(1). \end{aligned} \tag{113}$$

Since $\mathbb{E}\left(H_1\overline{\mu}_i^I - H_2\overline{\mu}_i^{II} - H_2 r_i\xi_i\right)^2 = \Omega_{22}/\Omega_2 \leq 1$, by M and Lemma B6 of Imbens and Newey (2002), we have

$$\left|\widehat{\Omega}_{22}/\Omega_2 - n^{-1}\sum_i \left(H_1\overline{\mu}_i^I - H_2\overline{\mu}_i^{II} - H_2 r_i\xi_i\right)^2\right| = o_P(1). \tag{114}$$

By LLN, we have

$$\left|n^{-1}\sum_i \left(H_1\overline{\mu}_i^I - H_2\overline{\mu}_i^{II} - H_2 r_i\xi_i\right)^2 - \Omega_{22}/\Omega_2\right| = o_P(1). \tag{115}$$

Therefore, by T, we obtain

$$\widehat{\Omega}_{22}/\Omega_2 - \Omega_{22}/\Omega_2 = o_P(1). \tag{116}$$

Combining results for $\widehat{\Omega}_{21}$ and $\widehat{\Omega}_{22}$, we have

$$\widehat{\Omega}_2/\Omega_2 - 1 \xrightarrow{p} 0. \tag{117}$$

□

C Notation

A_i : individual fixed effect

$A_1, \widehat{A}_1, A_2$: $A_1 = r^M(x)' \mathbb{E} r_i \overline{p}_i', \widehat{A}_1 = r^M(x)' \widehat{R}^{-1} \left(n^{-1} \sum_i r_i \widehat{\overline{p}}_i' \right), A_2 = r^M(x)$

$B, \widetilde{B}, \widehat{B}, B^X$: $(\beta_1, ..., \beta_n)', \left(\widetilde{\beta}_1, ..., \widetilde{\beta}_n\right)', \left(\widehat{\beta}_1, ..., \widehat{\beta}_n\right)', (\beta(X_1), ..., \beta(X_n))'$

d_X : dimension of X_{it}

d_1 : series approx rate for $V(x, z, w)$

d_2 : series approx rate for $G(s)$

d_3 : series approx rate for $\beta(x)$

F : $\Omega_2^{-1/2}$

$G(S), \widehat{G}(S)$: $\mathbb{E}[Y|X, V, W], \ p^K(S)' \widehat{\alpha}^K$

$H_1, \widehat{H}_1, H_2, \widehat{H}_2$: $H_1 = FA_1P^{-1}, \widehat{H}_1 = F\widehat{A}_1\widehat{P}^{-1}, H_2 = FA_2, \widehat{H}_2 = FA_2\widehat{R}^{-1}$

K : degree of basis functions $p^K(\cdot)$ used to estimate G

K_1 : degree of $p^{K_1}(\cdot)$, a component of $p^K(\cdot)$ and $\overline{p}^K(\cdot)$

L : degree of basis functions $q(\cdot)$ used to estimate V

M : degree of basis functions $r(\cdot)$ used to estimate $\beta(x)$

$p^K(s)$: $x \otimes p^{K_1}(v, w)$ for $s = (x, v, w)$, a $DK_1 \times 1$ vector

$\overline{p}^K(s)$: $I_D \otimes p^{K_1}(v, w)$, a $DK_1 \times D$ matrix

$p^{K_1}(v, w)$: component basis function of (v, w)

$q_i, p_i, \widehat{p}_i, \overline{p}_i, \widehat{\overline{p}}_i, r_i$: $q^L(X_i, Z_i, W_i), p^K(s_i), p^K(\widehat{s}_i), \overline{p}^K(s_i), \overline{p}^K(\widehat{s}_i), r^M(X_i)$

$$
\begin{aligned}
p, \overline{p}, \widehat{p}, \widehat{\overline{p}} &: (p_1, ..., p_n)^{'}, (\overline{p}_1, ..., \overline{p}_n)^{'}, (\widehat{p}_1, ..., \widehat{p}_n)^{'}, \left(\widehat{\overline{p}}_1, ..., \widehat{\overline{p}}_n\right)^{'} \\
q, r &: (q_1, ..., q_n)^{'}, (r_1, ..., r_n)^{'} \\
P, \widetilde{P}, \widehat{P} &: \mathbb{E} p_i p_i^{'},\ n^{-1} \sum p_i p_i^{'},\ n^{-1} \sum \widehat{p}_i \widehat{p}_i^{'} \\
Q, \widehat{Q} &: \mathbb{E} q_i q_i^{'},\ n^{-1} \sum q_i q_i^{'} \\
R, \widehat{R} &: \mathbb{E} r_i r_i^{'},\ n^{-1} \sum r_i r_i^{'} \\
s, S &: (x, v, w), (X, V, W) \\
U &: \text{random shock per period} \\
V &: F_{X|Z,W} \text{ control function for } U \\
W &: \text{sufficient statistic for } A \\
X &: \text{regressors for } Y, \text{ e.g. labor, capital} \\
Y_{it}, y &: \text{outcome variable e.g. value-added output, } y = (Y_1, ..., Y_n)^{'} \\
Z &: \text{instruments for } X, \text{ e.g. interest rate} \\
\mathcal{X}, \mathcal{Z}, \mathcal{W}, \mathcal{V}, \mathcal{S} &: \text{the support of } X, Z, W, V, S \\
s, x, z, w &: \text{realization of random variables} \\
X_{it}, Z_{it} &: \text{random vectors} \\
\mathbf{X}_i, \mathbf{Z}_i &: \text{random matrix } (X_{i1}, ..., X_{iT})^{'}, (Z_{i1}, ..., Z_{iT})^{'} \\
\alpha^K, \widehat{\alpha}^K &: \text{series approx coefficient for } G(s),\ \widehat{P}^{-1} \widehat{p}^{'} y / n \\
\beta_{it} &: \text{random coefficients} \\
\overline{\beta} &: \mathbb{E} \beta_{it} \\
\beta(x) &: \mathbb{E}\left[\beta_{it} | X_{it} = x\right] \\
\beta(v, w) &: \mathbb{E}\left[\beta_{it} | V_{it} = v, W_i = w\right] \\
\beta_v(v, w) &: \partial \beta(v, w) / \partial v \\
\beta_i, \widetilde{\beta}_i, \widehat{\beta}_i &: \beta(V_i, W_i), \beta\left(\widehat{V}_i, W_i\right), \widehat{\beta}\left(\widehat{V}_i, W_i\right) \\
\delta_{0t} &: \mathbb{E}\left[d_t(U_{2,it})\right] \\
\gamma^L(\cdot) &: \text{series approx coefficient for } V(x, z, w) \\
\eta^M &: \text{series approx coefficient for } \beta(x) \\
\lambda &: \text{eigenvalue of a matrix} \\
\text{psd, pd} &: \text{positive semi-definite, positive definite} \\
\overline{\mu}_i^I, \overline{\mu}_i^{II} &: \mathbb{E}\left[G_v(S_j) \tau^{'}(V_j) p_j q_j^{'} q_i v_{ji} \middle| \mathcal{I}_i\right], \mathbb{E}\left[\beta_v(V_j, W_j) r_j q_j^{'} q_i v_{ji} \middle| \mathcal{I}_i\right]
\end{aligned}
$$

$$
\begin{aligned}
\Omega_1 &: \overline{p}^K(s)' P^{-1}(\Sigma+\Sigma_1) P^{-1}\overline{p}^K(s) \\
\Sigma, \Sigma_1 &: \mathbb{E} p_i p_i' u_i^2,\ \mathbb{E}\overline{\mu}_i^I \overline{\mu}_i^{I'} \\
u_i, \widehat{u}_i &: Y_i - G(S_i),\ Y_i - \widehat{G}\left(\widehat{S}_i\right) \\
v_{ji} &: \mathbb{1}\{x_i \le x_j\} - F(x_j | z_i, w_i) \\
\widehat{\Omega}_1 &: \overline{p}^K(s)' \widehat{P}^{-1}\left(\widehat{\Sigma}+\widehat{\Sigma}_1\right) \widehat{P}^{-1}\overline{p}^K(s) \\
\widehat{\Sigma}, \widehat{\Sigma}_1 &: n^{-1}\sum_i \widehat{p}_i \widehat{p}_i' \widehat{u}_i^2, n^{-1}\sum_i \widehat{\mu}_i^I \widehat{\mu}_i^{I'} \\
\widehat{\overline{\mu}}_i^I, \widehat{\overline{\mu}}_i^{II} &: n^{-1}\sum_j \widehat{G}_v\left(\widehat{S}_j\right) \widehat{p}_j q_j' \widehat{Q}^- q_i \widehat{v}_{ji},\ n^{-1}\sum_j \widehat{\beta}_v\left(\widehat{S}_j\right) r_j q_j' \widehat{Q}^- q_i \widehat{v}_{ji} \\
\widehat{v}_{ji} &: \mathbb{1}\{x_i \le x_j\} - \widehat{F}(x_j | z_i, w_i). \\
\Omega_{21} &: \mathbb{E}\left(A_1 P^{-1} p_i u_i\right)\left(A_1 P^{-1} p_i u_i\right)' \\
\Omega_{22} &: \mathbb{E}\left[\left(A_1 P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i \xi_i\right)\right)\left(A_1 P^{-1}\overline{\mu}_i^I - A_2\left(\overline{\mu}_i^{II} + r_i \xi_i\right)\right)'\right] \\
\xi_i, \widehat{\xi}_i &: \beta(V_i, W_i) - \beta(X_i),\ \widehat{\beta}\left(\widehat{V}_i, W_i\right) - \widehat{\beta}(X_i) \\
\Omega_2 &: \Omega_{21} + \Omega_{22} \\
\widehat{\Omega}_{21} &: \widehat{A}_1 \widehat{P}^{-1}\left(n^{-1}\sum_i \widehat{p}_i \widehat{p}_i' \widehat{u}_i^2\right) \widehat{P}^{-1}\widehat{A}_1' \\
\widehat{\Omega}_{22} &: n^{-1}\sum_i \left(\widehat{A}_1 \widehat{P}^{-1}\widehat{\overline{\mu}}_i^I - \widehat{A}_2\left(\widehat{\overline{\mu}}_i^{II} + r_i \widehat{\xi}_i\right)\right)\left(\widehat{A}_1 \widehat{P}^{-1}\widehat{\overline{\mu}}_i^I - \widehat{A}_2\left(\widehat{\overline{\mu}}_i^{II} + r_i \widehat{\xi}_i\right)\right)'
\end{aligned}
$$

In the proofs :

CM : Conditional Markov Inequality

CS : Cauchy–Schwarz Inequality

LLN : Law of Large Numbers

M : Markov Inequality

T : Triangle Inequality

D Additional Simulation and Empirical Results

To provide more evidence on how well the proposed method can estimate the APE $\overline{\beta}$, we compare the histogram of the estimated $\widehat{\overline{\beta}}_b^{(d)}$ against the simulated APE $\overline{\beta}_b^{(d)} = (NT)^{-1}\sum_{i,t}\beta_{it,b}^{(d)}$, where $\beta_{it,b}^{(d)}$ is the dth dimension of the it-specific β_{it} for the bth round

of simulation, across all B simulations. Figure 8 compares the distribution of $\widehat{\overline{\omega}}_b$ with $\overline{\omega}_b$ across those B simulations. It shows that the proposed method can capture the dispersion of the true $\overline{\omega}_b$ reasonably well. The distribution of $\widehat{\overline{\omega}}_b$ centers around $\mathbb{E}\omega_{it} = 25/8$, echoing the findings in Table 1. It is also worthwhile mentioning that the majority of $\widehat{\overline{\omega}}_b$ lies in $[2.95, 3.4]$, a short interval relative to the size of $\mathbb{E}\omega_{it}$. Note that the distribution of $\widehat{\overline{\omega}}_b$ appears to be slightly right-skewed across B simulations.

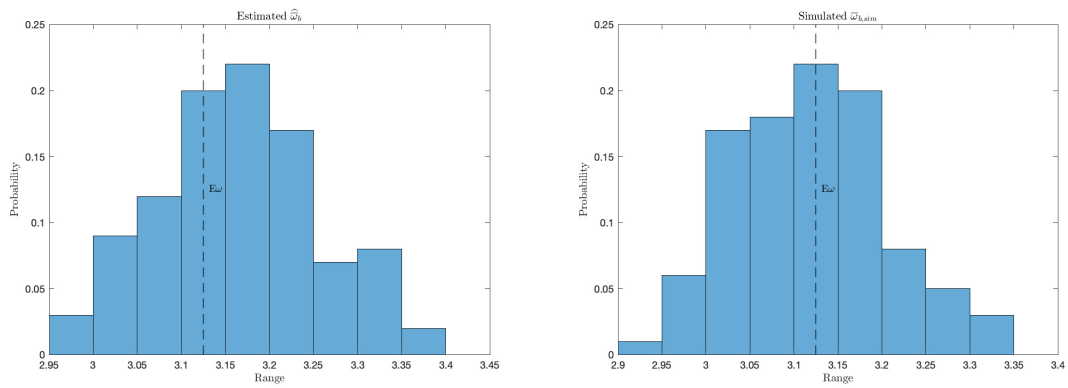

Figure 8: Histogram of $\widehat{\overline{\omega}}_b$ and $\overline{\omega}_b$

We conduct the same comparison for β^K and β^L and present the results in Figure 9 and 10, respectively. The results are similar to that obtained for ω. Once again, the method can capture the distributional characteristics of the true APE well, with the estimated coefficients located in a tight interval centered around the true APE.

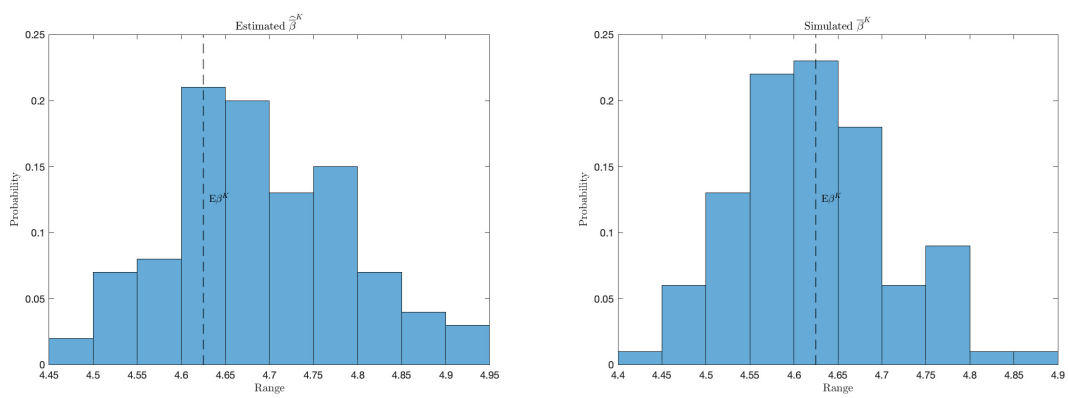

Figure 9: Histogram of $\widehat{\overline{\beta}}_b^K$ and $\overline{\beta}_b^K$

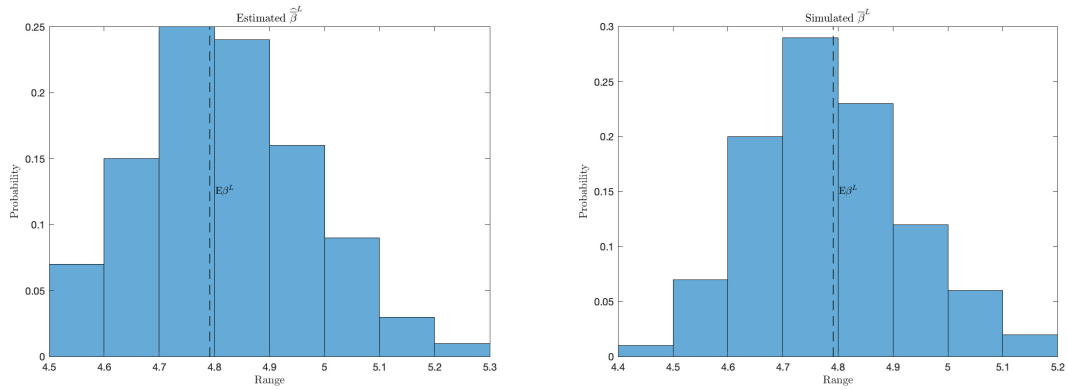

Figure 10: Histogram of $\widehat{\overline{\beta}}_b^L$ and $\overline{\beta}_b^L$

D.1 Robustness Checks

To show how robust the method is in estimating the APE, we conduct another set of exercises in this section. We evaluate the performance of the proposed method using both rMSE defined as $\sqrt{B^{-1}\sum_b \left\|\widehat{\overline{\beta}}_b - \overline{\beta}\right\|^2 / \left\|\overline{\beta}\right\|^2}$, and mean normed deviation (MND) defined as $B^{-1}\sum_b \left\|\widehat{\overline{\beta}}_b - \overline{\beta}\right\| / \left\|\overline{\beta}\right\|$.

First, we vary the size of N and T, and summarize the results in Table 3. As expected, a larger N is good for overall performance. We also find the proposed method benefit from the increase in T for each fixed N, possibly due to better controlling for the fixed effect A_i with more periods of data available for each individual.

Table 3: Performance under Varying N and T

	rMSE		MND	
	$N = 500$	$N = 1000$	$N = 500$	$N = 1000$
$T = 3$	0.0305	0.0298	0.0251	0.0242
$T = 5$	0.0241	0.0223	0.0206	0.0191

Second, we vary the order of the basis functions used to construct the series estimators, and present the results in Table 4. We find that increasing the orders of basis functions generally improves estimation accuracy. With that said, by using higher-order basis functions, one puts more pressure on the data because there are

more regressors in each step of estimation, which may explain why the improvement in performance from increasing the order of basis functions from two to three is significantly smaller than that from going from one to two. Motivated by the simulation result, we use a basis function with an order of two in the empirical illustration in the next section.

Table 4: Performance under Varying Orders of Basis Functions

Order of Basis Functions	rMSE	MND
1	0.0607	0.0562
2	0.0298	0.0242
3	0.0290	0.0237

Lastly, we examine how including ε_{it}, interpreted as measurement error or ex-post shock, into the model affects finite sample performance. Specifically, $\varepsilon_{it} \sim \mathcal{U}[-1/2, 1/2]$ is drawn independently from all other variables. Results are presented in Table 5. It is clear that adding ε_{it} negatively affects the performance of the proposed estimator, however the impact is mild. When ε_{it} is included, rMSE increases from 0.0298 to 0.0391 and MND rises from 0.0242 to 0.0318. The magnitude in the change in performance is small, showing that the proposed method is robust to the inclusion of measurement error.

Table 5: Performance with and without Ex-Post Shock

Ex-Post Shock?	rMSE	MND
No	0.0298	0.0242
Yes	0.0391	0.0318

E Additional Empirical Results

We show the distribution of $\widehat{\beta}(X_{it})$ for each sector in Figure 11. Note that by definition

$$\beta(x) = \mathbb{E}[\beta_{it} | X_{it} = x] = \mathbb{E}[\beta(V_{it}, W_i) | X_{it} = x].$$

Thus, $\beta(X_{it})$ can be considered as the average of $\beta(V_{it}, W_i)$ over individual firms with the same observable characteristics $X_{it} = x$. Consequently, the histogram of $\widehat{\beta}(X_{it})$

will be more concentrated around the mean than $\widehat{\beta}\left(\widehat{V}_{it}, W_i\right)$, as can be seen in Figure 11.

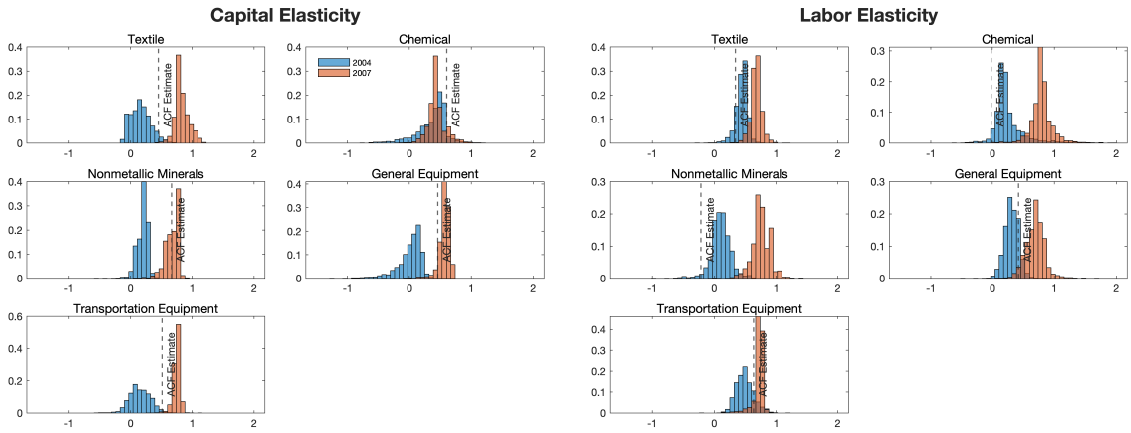

Figure 11: Distribution of $\widehat{\beta}^K(X_{it})$ and $\widehat{\beta}^L(X_{it})$