Semiparametric Panel Model with Group Heterogeneity

Peng Shao[1]

Abstract

This paper studies a semiparametric partially linear panel model with time-varying group-level effects. As a critical feature, the group memberships are unobserved but time-invariant. The linear coefficients estimator is shown asymptotically normal for inference. For production function estimation, the paper also considers a two-step problem; the objective (second-step) parameter is identified by moments, conditional on the partially linear model's potentially infinite-dimensional parameters. The paper proposes a second-step estimator and shows that it is consistent. The two analyses generically connect to the control function problem under the presence of time-varying heterogeneity for panel models. With the two-step solution, the paper extends the proxy variable method, designed for the simultaneity problem with estimating the firm's production function, by allowing cross-correlation in firms' productivity. As an empirical application, I consider four Chilean manufacturing sectors from 1987 to 1996. After accounting for cross-correlated productivity, I find larger productivity effects on output growth and more heterogeneous productivity among firms.

[1]Auburn University, Miller Hall, 402 W Thach Concourse, Auburn, AL 19104, pshao@sas.upenn.edu. I am deeply grateful to Frank Schorfheide, Xu Cheng, and Amit Gandhi for their support, suggestions, and insights. Also, I am grateful of Karun Adusumilli and Yuan Liao for their comments. I want to mention Frank DiTraglia and Wayne Gao for sharing their advice on my presentation. And I want to thank David Rivers and Amil Petrin for sharing their data sets with me. Lastly, I want to thank the participants at the UPenn Econometrics Lunch and Seminar for their comments.

1 Introduction

Econometricians often use panel data to control for economic agents' unobserved heterogeneity. The standard fixed effects model captures only time-invariant heterogeneity, but heterogeneity is plausibly time-varying in many applications. Furthermore, their dynamic processes can be governed by different time-varying trends, unknown to the econometrician.

For example, a firm's productivity is time-varying but unobserved to the econometrician. Economic policies to increase competition can lead to disparate outcomes between the productivity growth of efficient firms versus inefficient firms. Typically, it is nontrivial to discern efficient firms from their inefficient counterparts by just observing the data set.

A parsimonious but flexible solution is to treat agents' time-paths as time-varying group-level effects and leave each agent's group membership as unobserved. The dimension reduction of heterogeneity to the group-level helps to avoid the incidental parameter problem (associated with the fixed effects model) in the dynamic panel model with a short panel setup. Furthermore, the econometrician can recover the time-varying effects by pooling observations within the group per period and, consequently, treat time-paths' distributions as nonparametric. Finally, leaving group memberships as unobserved provides flexibility in allowing differences among agents in the panel.

This paper studies the semiparametric partially linear panel model with time-varying group-level effects. It consists of additive separable linear, nonparametric, and group-level effects components. In many economic applications, the partially linear model is used to control for observables' nonlinear effects captured by its nonparametric component. The econometrician often encounters nonparametric functions because the economic theory does not provide the parametric restrictions in applications. The nonlinear effects can be in the structural model or arise in the reduced-form as the control function. I expand this workhorse's tool-kit by introducing time-varying group-level effects with unobserved group memberships. In section 2, the paper provides economic examples for this partially linear model with group-level effects.

The paper applies the K-mean clustering idea to classify the group memberships and series estimation approach for the nonparametric function. Bonhomme and Manresa (2015) recently studied the linear model with the same group heterogeneity and used the K-mean approach. They coined this form of heterogeneity as the *grouped fixed effects*. Here, I study the grouped fixed effects in the context of the partially linear model. The series' strategy is to approximate the nonparametric function by a linear combination of parametric basis functions. This approach offers a computationally, simple method for nonparametric estimation.

The paper executes the group classification and the nonparametric estimation in a one-step approach. The asymptotic analysis is set up in the joint limit of $N, T \to \infty$ but have T as comparably small to N. This setup is consistent with the short wide panel. I provide sufficient conditions for the linear coefficient estimator as $\sqrt{NT}$-consistent and asymptotic normal - with the associated consistent covariance estimator. Furthermore, I also show uniform consistency of the grouped fixed effects estimator, nonparametric estimator, and the group membership classifier. Finally, I also propose an information criterion to determine the number of groups, and the sufficient conditions for its consistency are provided.

Subsequently, the paper considers a method of moments problem, conditioning on the

partially linear model's parameters, and proposes a second-step estimator, based on the sample moments conditioning on the partially linear model's estimators. This second-step estimates firms' production functions under the simultaneity issue and generalises the extensively used proxy variable method in the production function literature - see Ackerberg, Caves, and Frazer (2015). The proxy variable method assumes firms' productivity as mutually mean independent. I extend the approach by modeling grouped fixed effects as firms' cross-correlated productivity. Finally, the paper provides sufficient conditions for the second-step estimator's consistency.

As an empirical application, the paper estimates four Chilean manufacturing industry production functions from 1987 to 1996. The period covers the Chilean economic growth years after the Pinochet economic reforms. Furthermore, the production function literature extensively uses the Chilean data set to study in estimating production functions, and the paper visits this data set to account for firms' cross-correlated productivity. After accounting of firms' correlated productivity, the paper finds more significant productivity dispersion among firms, and productivity is generally more responsible for output growth than before. For example, the difference between productivity distribution's 75th and 25th percentiles widen by at least 50% percent for the second largest sector, Metal. Also, productivity's effect on the Metal sector's output growth gained an 18% increase, after controlling for input changes. Furthermore, highly productive firms dominate the market share and tend to hold more capital stock.

The paper extends Bonhomme and Manresa (2015)'s linear model by including an additive nonparametric term. In many economic applications, the nonparametric term arises to control for unobservables, as discussed by Blundell and Powell (2000), or to account for nonlinear effects. Unlike the linear setup, the classification problem now involves a new approximation error from the nonparametric estimation of m. The approximation error vanishes as the econometrician uses an increasingly more complex approximating model with larger sample sizes. However, there is a feedback effect between the classification error and the estimation error of numerous coefficients growing with the sample size. The paper shows Bonhomme and Manresa (2015)'s classification results hold in the partially linear case if the econometrician is sufficiently conservative in controlling the approximating model's growth relative to the panel's number of periods.

As already mentioned, the paper's classification approach is closest to the K-mean application by Bonhomme and Manresa (2015) and Bonhomme, Lamadon, and Manresa (2017). Parametric and nonparametric finite mixture models are more traditional approaches to classification, but they require either the parametric or estimated density. The K-mean method avoids the need for a correctly specified or consistently estimated density. More recently, Su, Shi, and Philips (2016) introduce the classifier-LASSO estimator as an alternative to K-mean. Both the K-mean and classifier-LASSO base their asymptotics on the joint limit of $N, T \to \infty$.

The partially linear model also connects with the literature of semiparametric and linear models using interactive fixed effects. As already noted by Bonhomme and Manresa (2015), grouped fixed effects is a useful alternative to Bai (2009)'s interactive fixed effects when N is comparably larger than T, like in the short wide panel. In particular, the linear coefficient estimator may require bias correction by using interactive fixed effects. Freyberger (2018), Huang (2013), and Su and Jin (2012) consider the interactive fixed effects in the

semiparametric setup. Like Bai (2009), they assume a strong factor setup - every cross-section unit's unobservable effect is a linear combination of the same factors. Hence every unobserved effect is assumed to be on a global scale. While grouped fixed effects restrict group members to have the same time effect, the setup does not preclude effects confined to a local scale. The two methods complement each other, and the preference of use should be context-dependent. On final note, Bai and Ando (2016) extend interactive fixed effects model to be group-specific. So they allow localized effects, but their bias issue remains.

The two-step problem connects to the literature of semiparametric conditional moment problem and generated regressors. Chen, Linton, and Keilegom (2003)'s general setup roughly covers the problem here, and the application relates more to the specifics of Olley and Pakes (1995). However, Chen, Linton, and Keilegom (2003) restrict their discussion to the cross-section setup with full independence in verifying their conditions. I verify Chen, Linton, and Keilegom (2003)'s conditions under grouped cross-sectional dependency and weak time dependency in the panel.

The paper's production function application bases on the proxy variable approach covered in a series of papers - Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg, Caves, and Frazer (2015). The proxy variable identifies the firm specific productivity, and the introduction of grouped fixed effects captures the cross-correlation in firms' productivity. The introduction relaxes productivity's first-order Markov assumption, mutual mean independence, and the scalar unobservable assumption in the proxy variable's setup.

Compared to alternatives, the proxy variable method is a minimalist in data requirement and also is flexible on the assumptions of productivity's dynamic and the market structure. The paper's extension enhances the method while it preserves the method's niche. For interpretation, the group structure partitions firms into the spectrum of high vs. low mean productivity level.

On the last note, Kasahara, Schrimpf, and Suzuki (2017) also studies production function heterogeneity with classification. However, they classify based on the parametric finite mixture model and make full assumptions on the market structure and productivity's dynamic. Here, the paper avoids making these assumptions. More recently, Cheng, Schorfheide, and Shao (2019) apply multidimensional K-means clustering to estimate different output elasticity and mean-level productivity efficiently; when the firm's productivity is autoregressive. Here, I assume homogeneous output elasticity but allow a more general productivity process.

The rest of the paper is organized as follows. Section 2 presents the partially model, and section 3 covers the two-step estimator for the production function. After the two-step estimator, section 4 presents the Monte Carlo simulation results. Then section 5 presents the empirical application. Finally, section 6 provides the conclusion. All proofs and supplementary materials are provided in the appendix.

2 Partially Linear Model

2.1 Model and Estimation

The Setup and Notations

In this section, I set up the notation for partially linear model. Subsequently, I will explain

my estimation strategy. The partially linear semiparametric panel model is,

$$y_{it} = x'_{it}\theta^0 + m\left(z_{it}\right) + \alpha^0_{it} + \epsilon_{it}, \qquad i, = 1, ..., N, \qquad t = 1, ..., T, \tag{1}$$

where the variables $(y_{it}, x_{it}, z_{it}, \alpha_{it}, \epsilon_{it}) \in \mathbb{R} \times \mathcal{X} \times \mathcal{Z} \times \mathcal{A} \times \mathbb{R}$ ($\mathcal{Z} \subset \mathbb{R}^{d_1}$, $\mathcal{X} \subset \mathbb{R}^{d_2}$, and $\mathcal{A} \subset \mathbb{R}$), the unknown function $m : \mathcal{Z} \to \mathbb{R}$, and the parameter $\theta^0 \in \Theta$, with $\Theta \subset \mathbb{R}^{d_2}$.

There are G^0 fixed groups, and each unit belongs to a group. The unit's membership is indexed as $g^0_i (\in \Gamma_{G^0} := \{1, ..., G^0\})$ and its membership is time-invariant. Within a group g, all its members share the time trajectory α_{gt}.

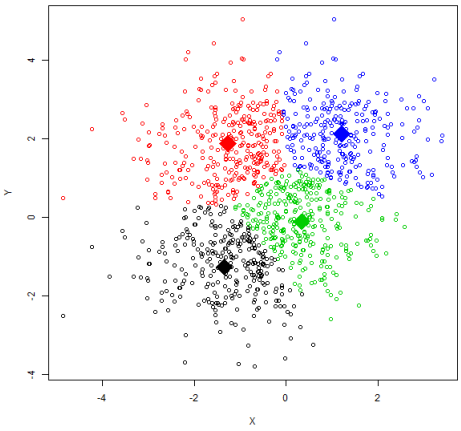
So,

$$\alpha^0_{it} = \begin{cases} \alpha^0_{1t} & g^0_i = 1 \\ \vdots & \vdots \\ \alpha^0_{G^0 t} & g^0_i = G^0. \end{cases} \tag{2}$$

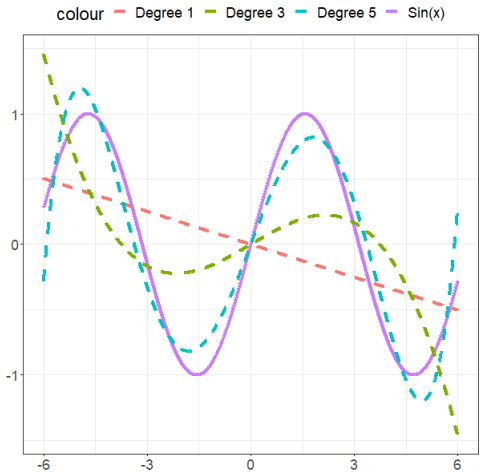
However, g^0_i is treated as unobserved from the data. From the panel data, the econometrician observes only (y_{it}, x_{it}, z_{it}). The (x_{it}, z_{it}) can be endogenous to $\alpha^0_{g^0_i t}$ but are sequentially exogenous to ϵ_{it}. For identification, I normalise $m\left(z'\right) = 0$ for some $z' \in \mathcal{Z}$ because m is nonparametric. Under appropriate rank conditions, the partially linear model's parameters are identified.[1] The model can serve to facilitate inference on θ^0 or to estimate θ, m, and α for a functional estimator's use, i.e. to act as generated regressors. Examples are provided in subsection 2.2.

Estimation

To estimate the parameters $\left(\theta^0, m, \alpha_{gt}, g^0_i\right)$, I simultaneously apply two econometric techniques: K-mean clustering to classify g^0_i and series approximation to estimate m.



K-Mean Clustering



Series Approximation: Power Series

The principle of K-mean is to detect the group structure, g^0_i, by partitioning the data around centroids. The left graph shows K-mean in action. Dotted observations are classified

[1] Consult Appendix-F for more details.

with color around four centroids, marked by diamond shapes. In the partially linear model's context, the centroids can be interpreted as the parameters, and the dotted observation's distance from its centroid is the residual ϵ_{it}.

The principle of series approximation is to use the sum of smooth functions to approximate an unknown function, m. The right graph shows this principle in action. The polynomials of x are trying to approximate $Sin\left(x\right)$, and the approximation error vanishes by increasing the polynomial degree. For smooth function m, the series approximation can be thought of as m's Taylor approximation.

Next I describe how to implement the two procedures jointly. Suppose the econometrician assumes the number of group is G and estimates the non-parametric $m(\cdot)$ with a vector of basis functions, $p^K(\cdot) = \left(p_1\left(\cdot\right), ..., p_K\left(\cdot\right)\right)$, where $p_s\left(\cdot\right) : \mathcal{Z} \to \mathbb{R}$, for $s = 1, ..., K$, and K is an integer. Furthermore, $\beta^K = \left(\beta_{1K}, ..., \beta_{KK}\right)'$ is the vector of coefficients for $p^K\left(z_{it}\right)$ to approximate m and $\beta_{sK} \in \mathcal{B}^K$. Popular example of p^K includes power series, Fourier series, and B-splines.

The group assignment $\gamma : \{1, ..., N\} \to \Gamma_G$, where $\gamma\left(i\right) = g_i$ and $\Gamma_G := \{1, ..., G\}$, denotes the collection of group membership parameters - with $\gamma^0 := \{g_i^0\}_{i=1}^N$. The partially linear model's estimator comes from minimizing the least-squared criterion:

$$\left(\hat{\theta}, \hat{\beta}^K, \hat{\alpha}, \hat{\gamma}\right) \in \underset{\theta \in \Theta, \beta^K \in \mathcal{B}^K, \alpha \in \mathcal{A}^{G \times T}, \gamma \in \Gamma_G^N}{\arg\min} \hat{Q}(\theta, \beta^K, \alpha, \gamma), \tag{3}$$

where $\hat{Q}\left(\theta, \beta^K, \alpha, \gamma\right) = \dfrac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta^K - p^K\left(z_{it}\right)'\beta^K - \alpha_{g_i t}\right)^2$. The estimator of m is $\hat{m}(z_{it}) = p^K(z_{it})'\hat{\beta}^K$.

The least-squared minimization problem is non-linear in γ. Next, I provide a simple local optimization algorithm to solve the problem. The optimization is local because it has the least-squared solution as a convergent point but its set of convergent points may not be singular.

Algorithm 1: Estimating θ, β^K, g_i, and α_{gt}

Initialize $\left\{\hat{g}_{i[0]}\right\}_{i=1}^N$.;
Using $\left\{\hat{g}_{i[0]}\right\}_{i=1}^N$, estimate $\hat{\theta}_{[0]}, \hat{\beta}_{[0]}^{K,j}$, and $\hat{\alpha}_{gt[0]}$ by minimizing the least-squared criterion;
while *convergence is not achieved on the kth iteration* **do**
  Using the kth iteration's $\hat{\theta}_{[k]}, \hat{\beta}_{[k]}^{K,j}$ and $\hat{\alpha}_{gt[k]}$, update $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^N$ to minimize the least squared criterion;
  Using $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^N$, estimate $\hat{\theta}_{[k+1]}, \hat{\beta}_{[k+1]}^{K,j}$ and $\hat{\alpha}_{gt[k+1]}$ by minimizing the least-squared criterion;
  Check for convergence of the modified least squared criterion;
end

The algorithm is convergent because it decreases the least-squared criterion at every step. For the global optimum, it is paramount to compute and compare local solutions using different initial arrangements of $\left\{\hat{g}_{i[0]}\right\}_{i=1}^N$.

Selection

In practice, the econometrician has to choose a G without knowing G^0. I assume the econometrician knows an upper bound $G_{\max}$ and a lower $G_{\min}$ for G^0, i.e. $G_{\min} \leqslant G^0 \leqslant G_{\max}$. For panel models, the information criterion is a popular tool for selecting the latent structure's complexity - see Bai and Ng (2002) and Su, Shi, and Philips (2016). For my partially linear model, the Information Criterion function is

$$IC(G) = \hat{Q}_G + \nu G, \tag{4}$$

where $\hat{Q}_G$ is the minimized least-squared criterion from using G groups and for some positive constant ν. The information criterion estimate of G^0 is

$$\hat{G^0} \in \underset{G \in \{G_{\min}, \ldots, G_{\max}\}}{\arg\min} IC(G). \tag{5}$$

I provide guidance in choosing ν in subsection 2.4 and provide an explicit example in section 4. The paper also provides the information criterion's consistency result.

2.2 Applications

Here, the paper provides three examples for the partially linear model's application. The first two examples' interest is on the coefficient θ. However, Example 3 estimates the firm's production function and uses the partially linear model's nonparametric component m. I only briefly sketch its estimation procedure, but section 3 expands the discussion on the production function estimation. Section 4 provides the Monte Carlo simulation covering Example 1 and Example 3. Section 5 covers an empirical application based on a more developed version of Example 3.

Example 1: (Income Growth vs. Inequality Growth) Banerjee and Duflo (2003) studies whether an increase in income inequality overall promotes or hinders $\left(\theta_1^0\right)$ income growth in a panel of countries. In their equation (9), income growth is affected by both inequality growth and level. However, only inequality growth is assumed to have a linear effect, while the inequality level may have a non-linear effect. Here, I present the simplified version as,

$$\Delta income_{it} = \theta_1^0 \Delta Gini_{it} + \theta_2^0 \Delta income_{it-1} + m(Gini_{it}) + \alpha_{g_i^0 t}^0 + \epsilon_{it}, \tag{6}$$

where the inequality level is measured by the Gini coefficient. Different economic policies and political institutions have varying effects on income growth. In their setup, a time-invariant country fixed effects captures the net outcome of these effects. Here, I propose to model these net effects as time-varying grouped fixed effects because policies' effects may change over time. The groups can partition countries by their independence of the judiciary, and market-based vs. central planning spectrum. To manually define the right partition is conceptually difficult and so it is practical to estimate the memberships.

Example 2: (A Stylized Wage Regression) The interest is in the linear effect (θ^0) of an additional year of education on logged wage for a particular sector and the researcher uses

a balanced panel of workers. Workers have unobserved but time-varying skill sets affecting their marginal product of labor. Their skill sets may change over time, and I propose to model them as grouped fixed effects. The stylized model is

$$\log\left(wage_{it}\right) = \theta^0 educ_{it} + \alpha^0_{g_i^0 t} + u_{it}. \tag{7}$$

However, the workers self-select into the data set, and this fact induces sample selection concerns. Here, I show two selection rules that are independent of $\alpha^0_{g_i^0 t}$.

Selection rule 1: (Wealth Effects) The worker only attends work if he can afford to delegate home-keeping tasks and some other life maintenance activities. These activities cost the worker v_{it} monetary value and, hence, the worker is only in the sample because $f\left(Wealth_{it}\right) \geqslant v_{it}$, where f is strictly increasing. Then the reduced form regression is,

$$\log\left(wage_{it}\right) = \theta^0 educ_{it} + \alpha^0_{g_i^0 t} + m\left(Wealth_{it}\right) + \epsilon_{it}, \tag{8}$$

where $m\left(Wealth_{it}\right) = \mathbb{E}\left[u_{it} | f^{-1}\left(u_{it}\right) \geqslant Wealth_{it}\right]$ and $\epsilon_{it} = u_{it} - \mathbb{E}\left[u_{it} | f^{-1}\left(u_{it}\right) \geqslant Wealth_{it}\right]$.

Selection rule 2: (Occupational Choice) The worker has an outside option in another sector s which pays

$$\log\left(wage_{it}\right) = \theta^0 educ_{it} + \alpha^0_{g_i^0 t} + u_{it,s}.$$

Here, $\alpha^0_{g_i^0 t}$ can be thought of as transferable skills set while ϵ is sector-specific skills set. So the worker is only in the sample because of $u_{it} \geqslant u_{it,s}$. Suppose there is a variable z_{it} that can act as a proxy of $u_{it,s}$, i.e. $u_{it,s} = f\left(z_{it}\right)$ where f is strictly increasing. Then the selection rule can be reduced to $f^{-1}\left(u_{it}\right) \geqslant z_{it}$ and provides the reduced form regression,

$$\log\left(wage_{it}\right) = \theta^0 educ_{it} + \alpha^0_{g_i^0 t} + m\left(z_{it}\right) + \epsilon_{it}, \tag{9}$$

where $m\left(z_{it}\right) = \mathbb{E}\left[u_{it} | f^{-1}\left(u_{it}\right) \geqslant z_{it}\right]$ and $\epsilon_{it} = u_{it} - \mathbb{E}\left[u_{it} | f^{-1}\left(u_{it}\right) \geqslant z_{it}\right]$.

The sample selection problem is an example of the general control function approach to control endogeneity. Generally, the partially linear model covers control function applications where the grouped fixed effects is not in the control function.

Example 3: The econometric objective is to estimate the output elasticity (β_k, β_l) of the log-linearized Cobb-Douglas production function,

$$y_{it} = \beta_l l_{it} + \beta_k k_{it} + \omega_{it} + \alpha^0_{g_i^0 t} + \epsilon_{it}, \tag{10}$$

where y_{it}, l_{it}, and k_{it} are logged output, labour, and capital, respectively. The sum $\omega_{it} + \alpha^0_{g_i^0 t} + \epsilon_{it}$ represents productivity but the firm can only learn $\omega_{it} + \alpha^0_{g_i^0 t}$ when it chooses inputs. ω_{it} is the firm's specific shock and independent over i. Furthermore, $\alpha^0_{g_i^0 t}$ is a productivity shock shared among the same typed firms - *it captures the cross-correlation in firms' productivity.* The firms are partitioned into groups by the productivity spectrum, e.g. high productivity type vs. low productivity type.

The regression has a simultaneity problem because the firm uses the information of $\omega_{it} + \alpha^0_{g_i^0 t}$ to choose (l_{it}, k_{it}) and the econometrician does not observe $\omega_{it} + \alpha^0_{g_i^0 t}$. Neglecting the simultaneity problem leads to "transmission bias" in the output elasticity estimates. Like in Olley and Pakes (1996) and Levinsohn and Petrin (2003), I assume there is a proxy variable v_{it} such that $\omega_{it} = h(l_{it}, k_{it}, v_{it})$, where h is an unknown function. The h function as independent of α is discussed in section 3 and section 3 also provides an example in the structural value-added setup. For brevity, I restrict my discussion to just the estimation procedure here.

The proxy variable turns the above regression into the reduced form,

$$y_{it} = m(l_{it}, k_{it}, v_{it}) + \alpha_{g_i^0 t} + \epsilon_{it}, \tag{11}$$

where $m(l_{it}, k_{it}, v_{it}) = \beta_l l_{it} + \beta_k k_{it} + h(l_{it}, k_{it}, v_{it})$. The nonparametric nature of h means the reduce form can't identify the output elasticity but isolates the productivity shocks $\alpha^0_{g_i^0 t} + \epsilon_{it}$. The first-step is to estimate the nonparametric function m with the partially linear model estimator. Then guessing output elasticity (β_k, β_l) provides an estimate of ω_{it} as $m(k_{it}, l_{it}, v_{it}) - \beta_k k_{it} - \beta_l l_{it} - \nu$[2]. Assuming ω_{it} is mean zero and independent over time then the output elasticity can be identified from lagged inputs' orthogonality condition to η_{it}, i.e.

$$\mathbb{E}\left[\begin{pmatrix} k_{it-1} \\ l_{it-1} \\ 1 \end{pmatrix} (m(k_{it}, l_{it}, v_{it}) - \beta_k k_{it} - \beta_l l_{it} - \nu)\right] = 0. \tag{12}$$

This formulation introduces a method of moments estimator for (β_k, β_l), conditioning on the first-step's estimates.

The section 3 generalises ω_{it} as a first-order Markov process and allows smooth Hicksian neutral technology. Furthermore, the section discusses the proxy variable assumption in more detail. Furthermore, the paper provides the general second-step method of moments estimator and sufficient conditions for its consistency.

2.3 Heuristics

It is instructive to motivate the theory's purpose heuristically before presenting it. The series estimation requires the researcher to choose the number of basis terms to approximate the unknown function. Bonhomme and Manresa (2015) provides asymptotic classification results for the linear model where there is an exact number of regressors. The series' approximation error of the unknown vanishes in the asymptotic by increasing the number of basis terms with the sample size. The theory has to extend the classification results by accounting for an increasing number of regressors.

To facilitate the discussion, I use an univariate semiparametric model:

$$y_{it} = m(z_{it}) + \alpha^0_{g_i^0 t} + \epsilon_{it}, \;\; g_i^0 \in \{1, 2\}, \tag{13}$$

[2] ν is normalization parameter to account for m's intercept as not separably identifiable from α_{gt}.

where ϵ_{it} is i.i.d., mean zero, and independent of α^0. The basis is the power series, $p^K(z_{it}) = \left(z_{it}, ..., z_{it}^K\right)'$. Basing classification on α^0 and parameter β_{kK}, the ith unit from group 1 is misclassified if

$$\sum_{t=1}^{T}\left(y_{it} - \sum_{k=1}^{K} z_{it}^k \beta_{kK} - \alpha_{2t}^0\right)^2 < \sum_{t=1}^{T}\left(y_{it} - \sum_{k=1}^{K} z_{it}^k \beta_{kK} - \alpha_{1t}^0\right)^2. \tag{14}$$

This inequality is equivalent to,

$$\frac{\sum_{t=1}^{T}\left(\alpha_{1t}^0 - \alpha_{2t}^0\right)^2}{T} < \text{“Approximation Error”} + o_p(1) + \sum_{k=1}^{K}\left(\beta_{kK}^0 - \beta_{kK}\right)\left[\frac{\sum_{t=1}^{T} z_{it}^k\left(\alpha_{2t} - \alpha_{1t}\right)}{T}\right],$$

where β_{kK}^0 is the target parameter of β_{kK}.

For identification purposes, I assume the left-hand side has a positive probability limit, i.e., the group effects are well-separated from each other. For consistent classification, the right-hand side needs to vanish asymptotically. The linear model does not have an approximation error term. Generally, the approximation error disappears by requiring the number of basis terms, K, to grow. But rapidly increasing the K terms can lead to an explosion of the third term even when $\left(\beta_{kK}^0 - \beta_{kK}\right)$ is small. This third term can be referred to as the “estimation error”.

The inequality also reveals a feedback back chain. When the estimation error is significant, classification error is likely to occur. Furthermore, it is intuitive to have a significant estimation error when the classification error is substantial. The standard series theory does control for the estimation error's magnitude, but it does not account for this feedback effect.

I propose to address this feedback effect by conservatively choosing K relative to T. This strategy allows me to mitigate the estimation error's impact. As T grows large, K is permitted to increase for the approximation error to vanish at the asymptotic.

The presented theory is to ensure asymptotic classification holds in a “worst-case” scenario. The inequality also suggests the above concern can be a second-order issue if the group effects are significantly well-separated from each other. However, the theory requires them as well-separated and does not assume a lower positive bound on the separation. Furthermore, the approximation error may be negligible in practice with just finitely many K terms when the unknown function is sufficiently smooth.

Besides the classification result, I provide sufficient conditions to derive $\hat{\theta}$ as asymptotically normal with its consistent covariance estimator. This result provides a central limit theorem when observations are weakly dependent over time and dependent over the cross-section via α^0. Consequently, I also derive the estimator of $\alpha_{g_i^0 t}$ and m as uniformly consistent to help the subsequent two-step estimator analysis. Finally, I provide sufficient conditions for the information criterion to estimate G^0 consistently when it is unknown. Furthermore, section 4 provides an explicit example. However, analysing post-selection effects is not considered here.

2.4 Asymptotic Theory

This section first proves the consistency of $\hat{\theta}$ and $\hat{m}$, as presented in Theorem 1. Theorem 2 proves the consistency of $\hat{G}$ when the information criterion's penalty satisfies appropriate

conditions. Then Theorem 3 shows classification error disappears asymptotically, i.e. $\hat{g}_i$ is uniformly consistent over i. Finally, Theorem 4 proves the asymptotic normality of $\hat{\theta}$ and uniform consistency of $\hat{\alpha}_{\hat{g}_i t}$.

The paper assumes the series $p^K(z_{it})$ to satisfy some high-level properties. For interpretation, high-level assumptions are elaborated for the power series and the B-spline series. The standard theory for the power series and the B-splines assume $\mathcal{Z}$ as a compact support. So the high-level assumptions are discussed with examples in the context of $\mathcal{Z}$ being compact. Chen (2007) provides other examples of series for the compact support. The discussion of the power series and B-splines can also apply to those series. All the assumptions are sequentially presented before theorems and progressively stronger to derive more demanding results. The provided asymptotic theory considers $N, T, K \to \infty$, but they grow at different rates. In particular, N is assumed to be significantly larger than T. Appendix-A collects all the asymptotic results.

Notation: Let $\|\cdot\|$ be the Euclidean norm, $\|f\|_{\infty,\mathcal{Z}} := \sup_{z\in\mathcal{Z}} \|f(z)\|$, $q_{it} = \left(x_{it}' \quad p^K(z_{it})'\right)'$, $\alpha = \left\{\{\alpha_{gt}\}_{t=1}^{\infty}\right\}_{g=1}^{G^0}$, and $x_{it} = (x_{it,1}, ..., x_{it,d_2})'$.

Assumption 1. *(Series approximation)*
There exist a constant $\mu > 0$ and the sequence $\left\{\left(\xi_K, \beta^{0,K}, \Pi_K\right)\right\}_{K=1}^{\infty}$, $\left(\xi_K, \beta^{0,K}, \Pi_K\right) \in \mathbb{R}_+ \times \mathcal{B}^K \times \mathbb{R}_+$, such that:

1. *$\|p^K\|_{\infty,\mathcal{Z}} \leqslant \xi_K$, and $\xi_K \to \infty$, as $K \to \infty$.*
2. *$\sup_{\beta\in\mathcal{B}^K} \max_{l=1,...,K} \|\beta_{lK}\| \leqslant \Pi_K$, where $\beta = (\beta_{1K}, ..., \beta_{KK})$, and Π_K is uniformly bounded away from 0.*
3. *$\left\|m - (p^K)'\beta^{0,K}\right\|_{\infty,\mathcal{Z}} = O(K^{-\mu})$ and $\beta^{0,K} \in \mathcal{B}^K$.*

Assumption 1.1 requires every finite-termed series to be bounded over the support $\mathcal{Z}$. In the following assumptions, the bound ξ_K should increase at a certain rate in proportion to the sample size. Newey (1997) provides a ξ_K as proportional to K for power series [3] and $\sqrt{K}$ for B-splines. Under Assumption 1.3, the series' approximation error of m over the entire support $\mathcal{Z}$ vanishes as the series' terms increase. The μ parameter captures the smoothness of the m function for both power series and B-splines. Newey (1997) shows $\mu = \delta^d/d_1$, where δ^d is number of m's continuous derivatives. Analogous Assumption 1.1 and Assumption 1.3 can be found in Newey (1997). Assumption 1.2 introduces a notation on the upper bound of $\beta^{0,K}$'s magnitude. The bound Π_K may increase over K subjected to the rates discussed later on. Π_K is constant for the power series if Assumption 1.3's approximation is also absolutely convergent in a neighborhood outside the unit ball. The examples include m being the sum of exponential functions, polynomials, and logarithms (when log takes values uniformly bounded away from zero). When Π_K is constant, the subsection's rates can drop the factor Π_K.

Assumption 2. *(Compactness)*
$\mathcal{A}$ and Θ are compact.

[3] The proportionality comes the orthogonal polynomial; which it spans the same linear space as the power series do.

The compactness of $\mathcal{A}$ rules out nonstationary α_{gt} process with its mean level growing over time. This same restriction is assumed by Bonhomme and Manresa (2015).

Assumption 3. *(Dependency and moment restrictions)*
There exist a constant $M > 0$ such that

1. $\sup_{i\in\{1,\dots,N\}} \mathbb{E}\left[\|x_{it}\|^4\right] \leqslant M$ *and* $\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}[\epsilon_{it}\epsilon_{is}x_{it}'x_{is}]\right| \leqslant M.$
2. $\mathbb{E}\left[u_{it}\right] = 0$ *and* $\sup_{i\in\{1,\dots,N\}} \mathbb{E}\left[u_{it}^4\right] \leqslant M.$
3. $\left|\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} Cov(\epsilon_{it}\epsilon_{jt}, \epsilon_{is}\epsilon_{js})\right| \leqslant M.$
4. $|\mathbb{E}\left[\epsilon_{is}\epsilon_{it}|z_{is}, z_{it}\right]| \leqslant M$ *and* $\mathbb{E}\left[\epsilon_{is}\epsilon_{jt}|z_{is}, z_{jt}\right] = 0$ *for any* $i \neq j$.

Assumption 3 restricts the dependency of ϵ_{it} on (x_{is}, z_{is}). Both x_{is} and z_{is} can be predetermined regressors. Overall, similar assumptions can be found in Bonhomme and Manresa (2015) and Bai (2009). However, Assumption 3.4 does not allow unconditional cross-correlation of ϵ_{it} and it also imposes bounded conditional heteroskedasticity and serial correlation. The bounded conditional heteroskedasticity assumption is standard in the series literature. If z_{it} is strictly exogenous then cross-correlation of ϵ_{it} can be restored by adopting Lee and Robinson (2016)'s approach.

Assumption 4. *(Rank Condition)*
Let $N^* := \lfloor \frac{N}{G} \rfloor - 1$ and $\mathcal{S} \subset \{1, ..., N\}$. If $\mathcal{S}$ has at least N^* units then

$$\mathbb{P}\left(\lambda_{\max}\left(\left[\frac{1}{TNN_S^2}\sum_{t=1}^{T}\sum_{i\in\mathcal{S}}\left(\sum_{j\in\mathcal{S}}(q_{it}-q_{jt})\right)\left(\sum_{j\in\mathcal{S}}(q_{it}-q_{jt})\right)'\right]^{-1}\right) < c\right) \underset{as\ N,T,K\to\infty}{\to} 1,$$

where $c > 0$, $\lambda_{\max}$ is the largest eigenvalue, and $N_S = \sum_{i=1}^{N}\{i \in S\}$.

Assumption 4 provides the rank condition to compute the least square estimator of $(\theta, \beta^K, \alpha)$, based on the estimated group memberships. For an arbitrary large group with at least N^* memberships, Assumption 4 requires sufficient cross-sectional variation of x_{it} and $p^K(z_{it})$ within the group. Hence, x_{it} and $p^K(z_{it})$ excludes constants.

Assumption 5. *(Rates and smoothness for consistency)*
As $N, T, K \to \infty$,

1. $K^{\frac{1}{2}-\mu}\xi_K^3 \amalg_K \to 0.$
2. $\frac{\xi_K^3\sqrt{K}\amalg_K}{\sqrt{N}} \to 0.$

3. $\frac{\xi_K^2}{T^{\frac{1}{4}}} \to 0$.

Assumption 5.1 controls effects from the vanishing approximation error. For power series and B-splines, Assumption 5.1 assumes m to be sufficiently smooth. Under the discussion of Assumption 1, when m has at least $4d_1$ continuous derivatives and $\amalg_K$ is bounded, e.g., real analytic functions, then Assumption 5.1 holds for both power series and B-splines. Assumption 5.2 and 5.3 restricts the series terms to asymptotically grow at a slower rate than N and T. Assumption 5.2 controls effects from the estimation error of the series' coefficients. However, Assumption 5.2's rate can be weakened to $\frac{\xi_K^3\sqrt{K}\amalg_K}{\sqrt{NT}} \to 0$ under the weak time dependency condition as specified in Assumption 9. In the cross-section setting, Newey (1997)'s semiparametric model and Qi (2000)'s partially linear model asks for $\frac{\xi_K\sqrt{K}}{\sqrt{N}} \to 0$. Assumption 5.2's rate is slower partly because parameters and group memberships are jointly estimated.

Assumption 5.3 is introduced to handle the unobserved group memberships. As mentioned in the outline, the rate expresses the caution of conservatively choosing K relative to T. Usually, the panel data literature allows the researcher to increase the basis' dimension by having a larger N. In contrast, Assumption 5.3 cautions against that behavior and leads to a more restrictive rate in the short wide panel setup.

Theorem 1. *(Consistency) Suppose Assumptions 1-5 hold and $G \geqslant G^0$, then 1) $\hat{\theta} \xrightarrow{P} \theta^0$, and 2) $\|m - \hat{m}\|_{\infty,\mathcal{Z}} \xrightarrow{P} 0$, as $N, T, K \to \infty$.*

Whenever the number of used groups is not smaller than the truth, Theorem 1 provides the consistency of $\hat{\theta}$ and the uniform consistency of $\hat{m}$. For just $\hat{\theta}$'s consistency, all Assumption 5's rates can be scaled down by ξ_K^2. However, having ξ_K^2 helps to show $\hat{m}$'s consistency. Moreover, subsequent results on classification and $\hat{\theta}$'s asymptotic normality depend on $\hat{m}$ being consistent. Improving Assumption 5's rates is an avenue for future work.

Assumption 6. *(Identifying Groups)*

1. *There exists a constant $c > 0$ such that,*

   (a) *when $g \neq g'$, then $\operatorname{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{g't}^0\right)^2 > c$, for a $c > 0$. This lower bound c applies to all pairs of g and g'.*

   (b) *for a real-valued process $\{h_t\}_{t=1}^{\infty}$ satisfying $\infty > \operatorname{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} h_t^2 > \frac{1}{2}c$, then*

$$\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} h_t^2 > \operatorname*{plim}_{N,T,K\to\infty}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} h_t q_{it}'\right)\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}q_{it}'\right)^{-1}\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}h_t\right).$$

2. *Let $N_g = \sum_{i=1}^{N}\{g_i^0 = g\}$. For any $g \in \{1, ..., G^0\}$, $\frac{N_g}{N} \underset{N\to\infty}{\to} \kappa_g > 0$.*

3. Assume $G = G^0$.

Assumption 6 provides the conditions to identify the groups. Assumption 6.1.a requires groups to be separately identified from their time-paths. Assumption 6.1.b is an identification assumption for the information criterion selection to avoid under-selecting. It basically says the differences between the groups effects should be far away from the regressors' spanned linear space. And Assumption 6.2 assumes each group's memberships is proportionally significant to the overall cross-section's sample size.

Corollary 1. *(Time-path consistency) Under Assumption 1-5, 6.1.a, 6.2, and* $G^0 \leqslant G$, *for any* $g \in \{1, ..., G^0\}$, *there exists a* $\hat{g}$ *such that* $\underset{N,T,K\to\infty}{\text{plim}} \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \hat{\alpha}_{\hat{g}t}\right)^2 = 0$. *Under Assumption 6.3,* $\hat{g}$ *is unique.*

With Corollary 1, each true group's α_{gt} time-path is matched asymptotically close to an estimated group's estimated time-path on average over time. This is weaker than uniform consistency but uniform consistency is achieved after additional stronger assumptions. Uniform consistency further provides consistency of $\hat{\alpha}_{\hat{g}_i t}$ at every i and t.

Assumption 7. *(Rates and smoothness for selection)*
As $N, T, K \to \infty$,

1. $\nu_T \to 0$.
2. $T^{\frac{1}{4}}\nu_T \to \infty$.
3. $\frac{T^{\frac{1}{4}}\xi_K \text{II}_K\sqrt{K}}{\sqrt{N}} = O_p(1)$.
4. $T^{\frac{1}{4}}K^{\frac{1}{2}-\mu}\xi_K \text{II}_K = O_p(1)$.
5. $G^0 \in \{\underline{G}, \overline{G}\}$.

Under Assumption 7.5, the information criterion minimizes over a set of specifications containing G^0. Under a large sample size, the logic behind the information criterion relies on over-specifying $G\left(> G^0\right)$ to yield negligible improvement and under-specifying $G\left(< G^0\right)$ leaves significant room for improvement in the least squared fit. To detect underfitting, Assumption 7.1 requires the penalty to vanish asymptotically. Moreover, to detect overfitting, Assumption 7.2 requires the penalty to vanish slowly at a rate dependent on only T. The T only dependency is set up under the assumption of N as comparably larger than T. In that environment, it is consistent with the information criterion literature to have the error rate as independent of N. For example, Bai and Ng (2002) have their rates as independent of N in the interactive factor setup when $\frac{\sqrt{T}}{N} \to 0$ - which is consistent with Assumption 7.3.

The information criterion's strategy for consistent selection also relies on the difference between the criteria, from over-specified and exactly specified, to vanish at a rate faster

than the penalty. Assumption 7.3 and 7.4 execute this task in combination. Furthermore, Assumption 7.3[4] and 7.4 act as Assumption 5.1 and 5.2 for the selection purpose, respectively.

Theorem 2. *(Selection) Suppose Assumption 1-4, 6.1, 6.2, and 7 hold, then* $\lim_{N,T,K\to\infty} \mathbb{P}(\widehat{G^0} = G^0) = 1$.

For the previous specific penalty, the selection is consistent for each λ. Hence, the data-driven choice of pre-specified λs is also consistent because the pre-specified set is finite; hence, the criterion is consistent for any λ of the set.

Theorem 2 shows the information criterion's estimate of G^0 is asymptotically consistent. Knowing the true number of groups is assumed for the subsequent theorems. However, the subsequent theorems do not account for the post-selection estimator.

Assumption 8. *(Tail-bounds)*

1. *There exist constants* $r_1 > 0$ *and* $r_2 > 0$ *such that,* $\mathbb{P}(|\epsilon_{it}| > m) \leqslant e^{1-\left(\frac{m}{r_1}\right)^{r_2}}$, *for all* i, t *and* $m > 0$. *For any* $i \in \{1, ..., N\}$,.
2. *For any* $g_l^0, g_k^0 \in \{1, ..., G^0\}$, $\mathbb{E}\left[\left(\alpha_{g_l^0 t}^0 - \alpha_{g_k^0 t}^0\right)\epsilon_{it}\right] = 0$.
3. *There exists constants* $r_3 > 0$ *and* $r_4 > 0$ *such that,* $\{\epsilon_{it}\}_{t=1}^{\infty}$, $\left\{\alpha_{g_j^0 t} - \alpha_{g_i^0 t}\right\}_{t=1}^{\infty}$ *and* $\left\{\left(\alpha_{g_j^0 t} - \alpha_{g_i^0 t}\right)\epsilon_{it}\right\}_{t=1}^{\infty}$ *are strongly mixing process, with mixing coefficient* $\rho_i(t)$, *and* $\sup_{i\in\{1,...,N\}} \rho_i(t) \leqslant e^{-r_3 t^{r_4}}$, *for any* $g_l^0, g_k^0 \in \{1, ..., G^0\}$.
4. *There exist constants* $M^* > 0$ *and* $\delta > 1$ *such that,*

$$\sup_{i\in\{1,...,N\}} T^{\delta}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\| \geqslant M^*\right) \to 0$$

and $\dfrac{N}{T^{\delta-1}} \to 0$, *as* $T, N \to \infty$.

Assumption 8.1, 8.2, and 8.3 assume tail bounds and weak dependency to control for classification error on the group membership estimate. Assumption 8.4 holds if x_{it}'s support is compact, and N is comparable to some power of T. Besides the comparable size of N and T, Assumption 8 is near identical to Bonhomme and Manresa (2015)'s Assumption 2 for their linear model case.

Theorem 3. *(Group Consistency) Let* $H_g^0 := \{i \mid g_i^0 = g\}$ *and* $\hat{H}_g := \{i \mid \hat{g}_i = g\}$. *When* $G^0 = G$ *and Assumption 1-6 and 8 hold, for any* $g \in \{1, ..., G\}$, *there exists* $g^0 \in \{1, ..., G^0\}$ *such that* $\mathbb{P}\left(\hat{H}_g = H_{g^0}^0\right) \to 1$, *as* $N, T, K \to \infty$.

[4]Just like Assumption 5.2, Assumption 7.3 can be weakened to $\dfrac{T^{\frac{1}{4}}\xi_K \text{II}_K\sqrt{K}}{\sqrt{NT}} = O_p(1)$ under weak time dependency.

When the exact total number of true groups is used, Theorem 3 says every estimated group asymptotically match to a true group in memberships. After some re-labeling of groups, Theorem 3 implies the uniform consistency of the group membership estimate, i.e $\mathbf{P}\left(\sup_{i\in\{1,...,N\}}|\hat{g}_i - g_i^0| > 1\right) = o_p(1)$. Furthermore, Theorem 3 also leads to $\left(\hat{\theta}, \hat{m}, \hat{\alpha}\right)$ as asymptotically equivalent to the Oracle estimator $\left(\tilde{\theta}, \tilde{\beta}^K, \tilde{\alpha}_{gt}\right)$. The Oracle estimator minimizes the least-squared criterion based on the true memberships.

The next assumption provides the additional conditions leading to uniform convergence of $\hat{\alpha}_{\hat{g}_i t}$, rate for $\hat{m}$, and the asymptotic normality of $\hat{\theta}$. To simplify presentation, I assume the moments, conditional of α, are identical within group. The proof uses the extended version, allowing heterogeneous conditional moments within the group. The extended version is provided in Appendix-A.

Assumption 9. *(Asymptotic Normality)*

1. *There exists a $\delta_m > 0$ such that $\max\left\{\left\|\theta - \theta^0\right\|, \left\|\beta^K - \beta^{0,K}\right\|\right\} < \delta_m$ implies $\beta^K \in \mathcal{B}^K$ and $\theta \in \Theta$. Furthermore, $\hat{\alpha}_{gt}$ is the interior solution.*

2. *For each $i \in \{1, ..., N\}$,*

   (a) *Conditional on α, $\{(x_{it}, z_{it}, \epsilon_{it})\}_{t=1}^{\infty}$ is independent over i.*

   (b) *Both conditional and unconditional on α, $(x_{it}, z_{it}, \epsilon_{it})$'s alpha mixing coefficient satisfies the uniform bound described in Assumption 8.3 up to a scale.*[5]

   (c) *ϵ_{it} is mean independent of x_{it}, z_{it}, and α.*

   (d) *the $(x_{it}, z_{it}, \epsilon_{it})$ process is stationary.*

3. *With some constant $C^{xp} > 0$, for any fixed K and $g \in \{1, ..., G^0\}$, the matrix*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(q_{it} - \mathbb{E}\left[q_{it}|\alpha\right]\right)\left(q_{it} - \mathbb{E}\left[q_{it}|\alpha\right]\right)'|\alpha, g_i^0 = g\right]' s$$

   smallest eigenvalue is bounded below by C^{xp}.

4. *Let $v(z_{it}) := (\mathbb{E}\left[x_{it,1}|z_{it}\right], ..., \mathbb{E}\left[x_{it,d_2}|z_{it}\right])$.*

   (a) *There exist sequences $\left\{\left\{\beta_{x,j}^K\right\}_{k=1}^{\infty}\right\}_{j=1,...,d_2}$ and constants $\{c_{vj}\}_{j=1,...,d_2}$ such that*

$$\sup_{j\in\{1,...,d_2\}}\left\|v_j - c_{vj} - \left(p^K\right)'\beta_{x,j}^K\right\|_{\infty,\mathcal{Z}} = O\left(K^{-\mu}\right),$$

   as $N, T, K \to \infty$, where μ is the constant in Assumption 1.

   (b) *There exists a positive constant M^m such that*

[5] It is not necessarily to assume Assumption 9.2 shares the same parametric values of r_3 and r_4 with Assumption 8.3. To avoid extra notation, I keep them as the same.

- $\sup_{t\in\{1,\ldots,T\}}\sup_{i\in\{1,\ldots,N\}}\mathbb{E}\left[m\left(z_{it}\right)^2\right]\leqslant M^m$,
- $\sup_{t\in\{1,\ldots,T\}}\sup_{i\in\{1,\ldots,N\}}\mathbb{E}\left[\left\|x_{it}\right\|^6\right]\leqslant M^m$,
- $\sup_{t\in\{1,\ldots,T\}}\sup_{i\in\{1,\ldots,N\}}\mathbb{E}\left[\left\|\left(x_{it}-\mathbb{E}\left[x_{it}\mid z_{it}\right]-\mathbb{E}\left[\mathbb{E}\left[x_{it}|\alpha\right]-\mathbb{E}\left[x_{it}|\alpha\right]|z_{it}\right]\right)\epsilon_{it}\right\|^5\right]\leqslant M^m$, *and*
- $\sup_{t\in\{1,\ldots,T\}}\sup_{i\in\{1,\ldots,N\}}\mathbb{E}\left[\left\|x_{it}-\mathbb{E}\left[x_{it}\mid z_{it}\right]-\mathbb{E}\left[\mathbb{E}\left[x_{it}|\alpha\right]-\mathbb{E}\left[x_{it}|\alpha\right]|z_{it}\right]\right\|^6|\{z_{is}\}_{s=1}^T,\alpha\right]\leqslant M^m$ *for any* α *and* $\{z_{is}\}_{s=1}^T$, *almost surely.*

(c) *There exist a sequence of constants* Π^x *such that* $\sup_{k\in\{1,\ldots,K\}}\left\|\beta^K_{x,j:k}\right\|\leqslant\Pi^x$ *and* $\frac{\sqrt{T}\Pi^x\sqrt{K}}{\sqrt{N}}\to 0$, *as* $N,T,K\to\infty$.

5. (a) *There exists a* $\delta'\in\left(0,\frac{1}{2}\right)$, *such that*
   - *i.* $\frac{T}{N^{\delta'}}\to 0$, *as* $N,T,K\to\infty$.
   - *ii.* $\frac{\xi_K^2\sqrt{K}}{N^{\frac{1}{2}-\delta'}\sqrt{T}}\to 0$, *as* $N,T,K\to\infty$.
   - *iii.* $\frac{N^{\delta'}\xi_K}{K^{\mu}}\to 0$, *as* $N,T,K\to\infty$.
   - *iv.* $\frac{\xi_K\sqrt{K}\Pi_K}{N^{\frac{1}{2}-\delta'}}\to 0$, *as* $N,T,K\to\infty$.

   (b) $\frac{T\sqrt{K}\xi^2\Pi_K}{\sqrt{N}}\to 0$, *as* $N,T,K\to\infty$.

   (c) $\frac{\sqrt{NT}\xi_K}{K^{\mu}}\to 0$, *as* $N,T,K\to\infty$, *where* μ *is the constant in Assumption 1.*

   (d) $\frac{K\xi_K}{T}\to 0$, *as* $N,T,K\to\infty$.

Assumption 9 provides sufficient conditions to derive the asymptotic distribution of $\hat{\theta}$ and uniform consistency of $\hat{\alpha}_{\hat{g}_it}$. The proof uses the usual least squared formula but this requires the solution of $\left(\hat{\theta},\hat{\beta}^K,\hat{\alpha}\right)$ to be in the interior. Theorem 1 and Assumption 9.1 provide $\left(\hat{\theta},\hat{\beta}^K\right)$ being in the interior with asymptotic probability one. However, until now, $\hat{\alpha}_{gt}$ is shown only to be mean-squared consistent over the sample path. This result is not enough to force it into the interior. However, verifying the solution as the interior is simple in practice.

Assumption 9.2 specifies weak dependency conditions. In the cross-section, it assumes α as the only source of cross-correlation for (x_{it},z_{it}) [6]. In time-series, the weak dependency is

[6]Potentially, the cross-section can allow weak dependency after conditioning on α. One possible extension is to use Lee and Robinson (2016)'s setup to model cross-sectional dependency. But, for simplicity, this weak dependency is not considered here.

described by mixing conditions. For example, (x_{it}, z_{it}) have the said alpha mixing properties if they are functions of independent processes with these alpha mixing properties.[7]

Assumption 9.3 strengthens the rank. The moments are defined conditionally because cross-sectional independence happens only conditional on α. But, conditional on α, the regressors q_{it} are not stationary. Hence, the condition bases on an average of over T.

From Assumption 9.4.a, the same series basis can uniformly approximate the conditional expectation of x_{it} - Qi (2000) uses a similar setup. Also, it assumes the conditional expectation of x_{it} is a homogeneous function over the cross-section. This restriction still allows x_{it} to be heterogeneous in expectation from the heterogeneity of distribution of z_{it} over the cross-section.

From Assumption 9.5, N is assumed as larger than T to ignore the incidental parameter problem of α_{gt}. So 9.5.b provides the rate for it to happen. Again, Assumption 9.5.c assumes m is sufficiently smooth such that scaling by $\sqrt{NT}$ still leaves the approximation error to be asymptotically negligible. Newey (1997) and Qi (2000) provide their analogous versions of Assumption 9.5.c to derive the asymptotic distribution. Assumption 9.5.a rates ensure the estimate of $\hat{\alpha}_{\hat{g}t}$ is uniformly consistent, over T, under the non-parametric estimation of m. However, Assumption 9.5.a is only relevant for the two-step problem and can be ignored for $\hat{\theta}$'s asymptotic normality. Assumption 9.5.d rate ensures the θ's asymptotic covariance matrix is convergent to its population analog under the non-parametric estimation of m. However, it is not needed to derive the consistency rate provided in the next theorem - thus, the two-step estimation can ignore this rate.

Theorem 4. *(Asymptotic Normality and Uniform Convergence)*
Under Assumption 1-6 and 8-9,

1. $\sqrt{NT}\left(\hat{\theta} - \theta^0\right) \Rightarrow N\left(0, \Sigma_\theta\right)$ *where*

$$a\ \Sigma_\theta = \left(\sum_{g=1}^{G^0} \kappa_g \psi_g^{xz}\right)^{-1} \psi^{x\epsilon} \left(\sum_{g=1}^{G^0} \kappa_g \psi_g^{xz}\right)^{-1},$$

$$b\ \psi_g^{xz} = \lim_{N\to\infty} \left[\frac{\sum_{i:g_i^0=g} \mathbb{E}\left[\left(x_{i1} - \psi^{xx\epsilon}\left(z_{i1}, \alpha\right)\right)\left(x_{i1} - \psi^{xx\epsilon}\left(z_{i1}, \alpha\right)\right)'\right]}{N_g}\right],$$

$$c\ \psi^{x\epsilon} = \lim_{N\to\infty} \frac{\sum_{i=1}^{N}\left[\sum_{t=1}^{\infty} \mathbb{E}\left[\left(x_{i1} - \psi^{xx\epsilon}\left(z_{i1}, \alpha\right)\right)\left(x_{it} - \psi^{xx\epsilon}\left(z_{it}, \alpha\right)\right)' \epsilon_{i1}\epsilon_{it}\right]\right]}{N},\ \textit{and}$$

$$d\ \psi^{xx\epsilon}\left(z_{it}, \alpha\right) = \mathbb{E}\left[x_{it} \mid z_{it}\right] + \mathbb{E}\left[x_{it} \mid \alpha\right] - \mathbb{E}\left[\mathbb{E}\left[x_{it}|\alpha\right]|z_{it}\right].$$

2. $\sup_{i\in\{1,\ldots,N\}, t\in\{1,\ldots,T\}} \left|\hat{\alpha}_{\hat{g}_i t} - \alpha^0_{g_i^0 t}\right| = o_p(1)$, *and*

3. $\left\|\hat{m} - m\right\|_{\infty,\mathcal{Z}} = O_p\left(\xi_K K^{-\mu}\right) + O_p\left(\xi_K^2 \frac{\sqrt{K}}{\sqrt{NT}}\right)$.

as $N, T, K \to \infty$.

[7] For reference, Andrews (1983) provides conditions to when a stationary autoregressive process is alpha mixing.

Theorem 4 provides the asymptotic normality for $\hat{\theta}$'s inference and α's estimates as uniformly consistent. Strengthening Corollary 1, Theorem 4 provides consistency of $\hat{\alpha}$ for everyone at every period. For the nonparametric estimate, the terms $O_p\left(\xi_K K^{-\mu}\right)$ and $O_p\left(\xi_K^2 \frac{\sqrt{K}}{\sqrt{NT}}\right)$ control the approximation and estimation errors, respectively. When moments are heterogeneous even within groups, the convergence rate has an extra term and the covariance matrix involves the group averaged $\mathbb{E}\left[x_{it} \mid \alpha\right]$ instead. The details are provided in the Appendix A.

The proof strategy of Theorem 4 relies on the asymptotic equivalence result implied by Theorem 3. Furthermore, Theorem 3's asymptotic equivalence implies that the classification problem leads to no efficiency loss in the limit. For θ's estimator, Robinson (1988) considers the semiparametric efficiency bound as the variance of θ's non-linear least squared (NLLS) estimator θ when m is a known parametric function identified by a finite-dimensional parameter γ_m. Then Robinson (1988) shows the double-residual semiparametric regression obtains this efficiency bound when $\mathbb{E}\left[x_{it}|z_{it}\right] = \dfrac{\partial m\left(z_{it};\gamma_m\right)}{\partial \gamma_m}$, almost surely.

The same exercise can be done here under no serial correlation, conditional homoskedasticity, x_{it} as strictly exogenous. In the presence of serial correlation or conditional heteroskedasticity, the NLLS's inefficiency is well-known. Moreover, including lags of x_{it} can also improve the NLLS estimator's efficiency when x_{it} is just sequentially exogenous. However, under those three conditions, the NLLS estimator is sensible a benchmark because it achieves the Gauss-Markov condition for its linear coefficient estimator.

Differencing the model by its group-level means turns the model into a simple partially linear model. By the Frisch-Waugh-Lovell theorem, the linear coefficient estimator obtained from applying NLLS on the demeaned model is identical to the version of NLLS applied to the original model. Then adapting Robinson (1988)'s observation and assuming m and its expectation are diferentiable in γ_m, the asymptotic semiparametric efficiency bound is

$$\sigma_\epsilon^2(\mathfrak{X}_1 - \mathfrak{X}_2\left[\mathfrak{X}_3\right]^{-1}\mathfrak{X}_2')^{-1}, \tag{15}$$

where $\sigma_\epsilon^2 = \mathbb{E}\left[\epsilon_{it}^2\right], \mathfrak{X}_1 = \mathbb{E}\left[\lim_{N\to\infty}\dfrac{\sum_{i=1}^N \breve{x}_{i1}\breve{x}_{i1}'}{N}\right], \mathfrak{X}_2 = \lim_{N\to\infty}\dfrac{\sum_{i=1}^N \mathbb{E}\left[\breve{x}_{i1}\frac{\partial}{\partial\gamma_m}\breve{m}\left(z_{i1};\gamma_m\right)'\right]}{N}, \mathfrak{X}_3 = \lim_{N\to\infty}\dfrac{\sum_{i=1}^N \mathbb{E}\left[\frac{\partial}{\partial\gamma_m}\breve{m}\left(z_{i1};\gamma_m\right)\frac{\partial}{\partial\gamma_m}\breve{m}\left(z_{i1};\gamma_m\right)'\right]}{N}$, $\breve{x}_{it} = x_{it} - \mathbb{E}\left[x_{it}|\alpha\right]$, and $\breve{m}\left(z_{it};\gamma_m\right) = m\left(z_{it};\gamma_m\right) - \mathbb{E}\left[m\left(z_{it};\gamma_m\right)|\alpha\right]$. Moreover,

$$\Sigma_\theta = \sigma_\epsilon^2\left[\lim_{N\to\infty}\frac{\sum_{i=1}^N \mathbb{E}\left[\left(x_{i1} - \psi^{xx\epsilon}\left(z_{i1},\alpha\right)\right)\left(x_{i1} - \psi^{xx\epsilon}\left(z_{i1},\alpha\right)\right)'\right]}{N}\right]^{-1} \tag{16}$$

under those three conditions. Now it is apparent that Σ_θ obtains the semiparametric bound when $\mathbb{E}\left[x_{it}|z_{it}\right] = \dfrac{\partial m\left(z_{it};\gamma_m\right)}{\partial\gamma_m}$, as in Robinson (1988). The efficiency bound argument easily extends to the case of heterogeneous moments within-group, as described in the extended Assumption 9.

Now I propose an estimator for Σ_θ under all the previous assumptions. I construct the sample analogs $\widehat{\psi_g^{xz}} = \frac{\sum_{i:\hat{g}_i=g}\sum_{t=1}^{T}\tilde{x}_{it}\tilde{x}'_{it}}{\hat{N}_g T}$ and $\widehat{\psi^{x\epsilon}} = \frac{\sum_i^N\left[\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{x}_{it}\tilde{x}'_{is}\hat{\epsilon}_{it}\hat{\epsilon}_{is}\right]}{NT}$ - where $\hat{\epsilon}_{it}$ denotes the partially linear model's residuals. Recall, $p^K(z_{it})$ denotes the basis. $\tilde{x}_{it,d}$ is the residual from the least-squared projection of $x_{it,d} - \frac{\sum_{i:\hat{g}_j=\hat{g}_i} x_{jt,d}}{\hat{N}_{\hat{g}_i}}$ onto $p^K(z_{it}) - \frac{\sum_{j:\hat{g}_j=\hat{g}_i} p^K(z_{jt})}{\hat{N}_{\hat{g}_i}}$, with $\hat{N}_g = \sum_{i=1}^{N}\{\hat{g}_i = g\}$.

Corollary 2. *(Covariance Estimator)*
Under Assumption 1-6, and 8-9, $\hat{\Sigma}_\theta = \left(\sum_{g=1}^{G^0}\frac{\hat{N}_g}{N}\hat{\psi}_g^{xz}\right)^{-1}\hat{\psi}^{x\epsilon}\left(\sum_{g=1}^{G^0}\frac{\hat{N}_g}{N}\hat{\psi}_g^{xz}\right)^{-1}$ *is a consistent estimator of* Σ_θ.

The covariance formula is the version of Arellano (1987)'s within-group estimator after accounting for the non-parametric estimation of $\hat{m}$. In their appendix, Bonhomme and Manresa (2015) also considers the within-group estimator for the linear model. Within the large N and T framework, Hansen (2007) shows the within-group estimator as consistent for the linear model. I extend this consistency result into the partially linear case. In section 4, I assess the covariance estimator's performance in constructing confidence intervals for a dynamic panel model with heteroskedasticity.

3 Two-Step Estimator

Here, I study a method of moments problem, conditioning on the partially linear model's estimated parameters. The goal is to set up a two-step estimator for the firm's production function. The first subsection expands on section two's discussion about the production function estimation and sets up a moment criterion example. The next section discusses the identification in more detail and provides a brief literature review of production function estimation near the end. Finally, I present a generic method of moments problem, conditioning on the partially linear model's estimated parameters. This general setup covers the production function estimation's criteria built from differing choices of moments. The asymptotic theory subsection shows that the general setup's two-step estimator is consistent.

3.1 Production Function

The objective is to estimate a parametric production function under the simultaneity problem - the firm bases its input choices on productivity, which is unobserved by the econometrician. Neglecting the simultaneity problem induces biases in the estimates - usually known as the "transmission bias". First, I generalise section two's example to allow smooth Hicksian production technology and ω_{it} as a first-order Markov process. The next subsection's conditional method of moments problem analyses this application under a general combination of moments.

At the high-level, my setup generalises the proxy variable approach on estimating the firm's production function. The generalisation introduces four additional features. The proxy variable approach assumes that different firms' productivity processes are first-order Markov and independent of each other. Furthermore, the Markov transition function is identical. I generalise by *introducing cross-correlation in firms' productivity* (1) and *relaxing the first-order Markov assumption* (2). It is plausible to assume the cross-correlation is relevant in application because spillover effects happen from technological advancements. Finally, I *weaken the scalar unobservable assumption* (3) and *allow firms' productivity transition dynamics to differ* (4).

After setting up the estimation procedure, I discuss my identifying assumptions. For readers only interested in the econometric setup, it is sufficient to read just the *Setup and Estimation* and the *Conditional Method of Moments* parts in this subsection.

Setup and Estimation

The Hicksian Neutral technology specification defines productivity to have the same effect on the marginal product of capital and labour. So the firm's production function is expressed as

$$Y_{it} = \exp\left(\epsilon_{it} + \alpha^0_{g_i^0 t} + \omega_{it}\right) F\left(K_{it}, L_{it}\right), \tag{17}$$

where Y_{it}, K_{it}, and L_{it} are output, capital, and labour, respectively. The production function can be log-linearised to form:

$$y_{it} = f\left(k_{it}, l_{it}; \tilde{\tau}\right) + \alpha^0_{g_i^0 t} + \omega_{it} + \epsilon_{it}, \tag{18}$$

where $y_{it} = \log\left(Y_{it}\right), k_{it} = \log\left(K_{it}\right)$, and $l_{it} = \log\left(L_{it}\right)$. The econometric objective is to estimate the parameter $\tilde{\tau}$ when the parametric form of f is known. For Cobb-Douglas case, $f\left(k_{it}, l_{it}\right) = \beta_k k_{it} + \beta_l l_{it}$ with $\tilde{\tau} = (\beta_k, \beta_l)$ and, for Constant Elasticity of Substitution case, $f\left(k_{it}, l_{it}\right) = \beta_1 \log\left(\exp\left(k_{it}\beta_2\right) + \exp\left(l_{it}\beta_2\right)\right)$ with $\tilde{\tau} = (\beta_1, \beta_2)$.

As before, I assume there is a variable υ_{it} to proxy ω_{it} after accounting for the firm's input choice of (k_{it}, l_{it}). That is $\omega_{it} = h\left(k_{it}, l_{it}, \upsilon_{it}\right)$, for some unknown function h. A popular choice of υ_{it} is the firm's intermediate material input in the recent applied literature. In the next part, I briefly discuss when the h function is independent of α^0. Furthermore, I provide an example of the h function that is independent of α^0 in the *Structural Examples* part.

The h function provides the reduced form regression,

$$y_{it} = m\left(k_{it}, l_{it}, \upsilon_{it}\right) + \alpha^0_{g_i^0 t} + \epsilon_{it}, \tag{19}$$

where the nonparametric function $m\left(k_{it}, l_{it}, \upsilon_{it}\right) = f\left(k_{it}, l_{it}; \tilde{\tau}\right) + h\left(k_{it}, l_{it}, \upsilon_{it}\right)$. In the absence of α^0, the reduced form regression is a standard first-step regression in the extensively used proxy variable's method to estimate the production function. As usual in the literature, h is treated as nonparametric and thus f is not separably identifiable from h. Then the reduced form regression is a semiparametric regression and can be estimated by the partially linear model's method in section 2. This estimation forms the first-step and separates ω_{it} from $\alpha^0_{g_i^0 t} + \epsilon_{it}$. Hence, this estimation step can also be referred as the filtering step.

Let $v_{it} = \omega_{it} - \mathbb{E}\left[\omega_{it} \mid \omega_{it-1}\right]$. For now, I assume the following moment conditions are valid:

$$\mathbb{E}\left[\begin{pmatrix} k_{it-1} \\ l_{it-1} \\ 1 \end{pmatrix} v_{it}\right] = 0 \text{ and } \mathbb{E}\left[\omega_{it}\right] = 0. \tag{20}$$

The first set of moments assumes the firm to forecast ω_{it} with just the information of ω_{it-1} when the firm chooses capital and labour inputs at the period $t-1$. For now, I assume $\tilde{\tau}$'s dimension is less than four for these moment conditions to sufficiently identify it. In the next part, assumptions are presented and more valid moments appear to identify $\tilde{\tau}$ when it has higher dimensions.

Next, I discuss how to construct sample moment analogues to estimate $\tilde{\tau}$. Based on m's identity, $\omega_{it} = m\left(k_{it}, l_{it}, v_{it}\right) - f\left(k_{it}, l_{it}; \tilde{\tau}^0\right)$[8]. Upon having the estimate $\hat{m}$, guessing $\tilde{\tau}^0 = \tilde{\tau}$ leads to a guess of

$$\hat{\eta}_{it}\left(\tau\right) = \hat{m}\left(k_{it}, l_{it}, v_{it}\right) - f\left(k_{it}, l_{it}; \tilde{\tau}\right) - \nu, \tag{21}$$

where $\tau = \left(\tilde{\tau}, \nu\right)$. The constant ν appears from m's intercept as not separately identified from $\alpha^0_{g_i^0 t}$ in the filtering stage.

To recover v_{it}, I estimate $\mathbb{E}\left[\omega_{it} \mid \omega_{it-1}\right]$ by minimizing the least-squared prediction error of $\hat{\omega}_{it}\left(\tau\right)$ based off the series of basis functions $b^L\left(\hat{\omega}_{it-1}\left(\tau\right)\right)$. With the estimated firm's Markov prediction as $\hat{R}\left(\hat{\omega}_{it-1}\left(\tau\right)\right) = \hat{\omega}_{it-1}\left(\tau\right)' \hat{r}^L\left(\tau\right)$ then

$$\hat{v}_{it}\left(\tau\right) := \hat{\omega}_{it}\left(\tau\right) - \hat{R}\left(\hat{\omega}_{it-1}\left(\tau\right)\right). \tag{22}$$

For sections 4 and 5, I use the Cobb-Douglas production function and, hence, $\tilde{\tau}$ has dimension two. At there, the second-step's General Method of Moment (GMM) criterion is

$$\hat{\mathfrak{M}}_{NT}\left(\tau\right)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{\sum_{t=2}^{T} l_{it-1}\hat{v}_{it}\left(\tau\right)}{T}\right)^2 + \left(\frac{\sum_{t=2}^{T} k_{it-1}\hat{v}_{it}\left(\tau\right)}{T}\right)^2 + \left(\frac{\sum_{t=1}^{T} \hat{\omega}_{it}\left(\tau\right)}{T}\right)^2 + \left(\frac{\sum_{t=2}^{T} \hat{v}_{it}\left(\tau\right)}{T}\right)^2\right].$$

Thus the estimator

$$\hat{\tau} \in \arg\min_{\tau} \hat{\mathfrak{M}}_{NT}\left(\tau\right).$$

For identification purposes, the criterion does not pool the moments over the cross-section. When cross-sectional units have non-identical distributions, pooling the cross-sectional moments can cause the criterion to have non-unique minima.

The assumptions and the relationship with the proxy variable method

Here, I present my assumptions about the firm's behaviour on the proxy variable. Then I compare them to the standard assumptions in the literature, as presented by Ackerberg, Caves, and Frazer (2015).

Assumptions:

[8] $\tilde{\tau}^0$ stands for the true value of $\tilde{\tau}$.

1. (Exclusion): v_{it} is neither capital nor labor.
2. (Scalar Unobservable): $v_{it} = \mathfrak{g}_t(k_{it}, l_{it}, \omega_{it})$.
3. (Strict Monotonicity): $\mathfrak{g}_t$ is strictly increasing in ω_{it}.
4. (Time invariance): $\mathfrak{g}_t(k_{it}, l_{it}, \omega_{it}) = \mathfrak{g}(k_{it}, l_{it}, \omega_{it})$.
5. (First-Order Markov): Let $\mathcal{I}_{it}$ be the firm's information set capturing all the firm's knowledge at the end of period t. $\mathbb{E}[\omega_{it} \mid \mathcal{I}_{it-1}] = \mathbb{E}[\omega_{it} \mid \omega_{it-1}]$. Furthermore, ω_{it} is a zero-mean process.
6. ("Surprise" shock): $\mathbb{E}[\epsilon_{it} \mid \mathcal{I}_{it}] = 0$.

It is instructive to first see how these assumptions sets up the estimation. The combination of assumptions 2, 3, and 4 imply $\mathfrak{g}$ as invertible with respect to ω_{it}, conditional on k_{it} and l_{it}. Then the unknown h is $\mathfrak{g}^{-1}(k_{it}, l_{it}, v_{it})$.

Assumption 5 verifies the provided moment conditions because k_{it-1} and l_{it-1} are in the firm's information set $\mathcal{I}_{it-1}$. Furthermore, Assumption 5 provides additional moment conditions,

$$\mathbb{E}\left[\begin{pmatrix} k_{it-s} \\ l_{it-s} \\ 1 \end{pmatrix} v_{it}\right] = 0 \text{ and } \mathbb{E}\left[\begin{pmatrix} k_{it-s} \\ l_{it-s} \\ 1 \end{pmatrix} (v_{it} + \epsilon_{it})\right] = 0, \text{ for } s \geqslant 1, \tag{23}$$

because the further lags are also in $\mathcal{I}_{it-1}$. The second set of moments can be constructed by using the sample analogue,

$$\widehat{v_{it} + \epsilon_{it}}(\tau) = y_{it} - \hat{m}(k_{it}, l_{it}, v_{it}) - f(k_{it}, l_{it}; \tilde{\tau}) - \hat{\nu} - \hat{R}(\hat{\omega}_{it-1}(\tau)) - \hat{\alpha}_{\hat{g}_i t}. \tag{24}$$

These additional moments would help to identify $\tilde{\tau}$ when its dimensional is greater than four. The section's last part sets up the notation for the general method of moments problem.

In absence of α^0 (i.e. $\alpha^0_{g_i^0 t} = 0$), the first, second, third, and fifth assumptions are standard in the proxy variable literature. In the sixth assumption, I interpret ϵ_{it} as the unpredictable productivity shock, as first suggested by Olley and Pakes (1996). The fourth assumption assumes $\alpha^0_{g_i^0 t}$ to completely capture changes in the macroeconomic environment. This consequence is more restrictive than the general proxy variable framework - I call this the time-invariant proxy variable. Section 7[9] discusses on how to handle $\mathfrak{g}_t$ with finitely many structural changes over time. However, the time-invariant setup is the often adopted specification in practice.[10] The pertinent observation is my setup nests the time-invariant proxy variable model.

Logged capital investment (Olley and Pakes (1996)) and logged intermediate material (Levinsohn and Petrin (2003)) are two popular choices of v_{it} in the production function literature. When v_{it} is intermediate material, the function h is the firm's conditional demand

[9] This is located after the conclusion section.

[10] Allowing time-varying $\mathfrak{g}_t$ requires in splitting the observations to estimate multiple nonparametric functions.

of intermediate material. When v_{it} is the capital investment, the function h is the firm's investment demand.

Assumption 2 says firm's demand function of v_{it} is constant over $\alpha^0_{g_i^0 t}$, after conditioning on $(k_{it}, l_{it}, \omega_{it})$. As an example, this Assumption 2 holds for intermediate material when the firm's capital and labor input choices define its production capacity. Then the firm uses intermediate material to fill up its production capacity. ω_{it} can be understood as productivity that scales up the firm's capacity while $\alpha^0_{g_i^0 t}$ does not.

Say, for instance, the firm produces twenty defected units of goods for every two hundred in production. However, the firm can only sell its non-defected units. With better training and quality control, the firm can reduce its defect rate down to five percent; then, this change is a productivity increase. However, the amount of intermediate material used to produce each unit remains unchanged, and then $\alpha^0_{g_i^0 t}$ captures this productivity increase. More discussions about h as constant over α^0 are provided in the *Structural Examples* part.

Under the presence of $\alpha^0_{g_i^0 t}$, the fifth assumption is a bit nuanced. There is an implicit assumption of ω_{it} as mean independent of α^0 conditional on ω_{it-1}. Relaxing this assumption is straightforward, but it is kept for simplicity. All forms of cross-correlation is to be absorbed by $\alpha^0_{g_i^0 t}$ and this leaves ω_{it} as independent over i.

The standard proxy variable model precludes dynamic cross-correlation in firms' productivity because of the fifth assumption, and $\alpha^0_{g_i^0 t} = 0$. It is to imagine the firm observing (at least partially) its competitors' productivity. So other firms' productivity information should be in the set $\mathcal{I}_{it}$. The fifth assumption says their information is not helpful to predict tomorrow's ω_{it+1} beyond knowing today's ω_{it}. For application, this means the firm's competitors can not independently innovate with positive spillover effects for the industry.

In the absence of $\alpha^0_{g_i^0 t}$, the Scalar Unobservable assumption with strict monotonicity predicts the firm to always increase its input level of v_{it} by a higher overall productivity level. This prediction is a reasonable assumption in the competitive market but can fail when the firm has market power. For example, if technological progress helps the firm to reduce its waste of intermediate material, then the firm with market power may want to cut its intermediate material purchases to raise profits. Then the strict monotonicity fails. With $\alpha^0_{g_i^0 t}$, the firm does not need to increases its purchases in a strict fashion with overall productivity.

Finally, $\alpha^0_{g_i^0 t}$ does not need to be a first-order Markov process. By including $\alpha^0_{g_i^0 t}$, the econometrician can be somewhat agnostic about the productivity process' order of persistence. Furthermore, the firms' productivity transition functions can now be different because $\alpha^0_{g_i^0 t}$ differs among firms. So my setup generalises the first order Markov setup used in the proxy variable approach.

Structural Examples

Here, I provide some worked out examples of $\mathfrak{g}$ as not dependent of $\alpha^0_{g_i^0 t}$.

Intermediate Material - v_{it} as logged intermediate material
The first example is in the setup of the structural value-added model, which is the frequently used example to justify the proxy variable assumptions - see Ackerberg, Caves, and Frazer

(2015) and Gandhi, Navarro, and Rivers (2017b). In that setup, F is the firm's "valued-added" production function but the firm has a gross production function described by the Leontief specification,

$$Y_{it} = \exp\left(\alpha^0_{g_i^0 t} + \epsilon_{it}\right) \min\{\mathcal{C}\left(M_{it}\right), \exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\}, \tag{25}$$

where M_{it} is the intermediate material and $\mathcal{C}$ is strictly increasing. But $\alpha^0_{g_i^0 t}$ is the same constant for every firm in the usual structural value-added model. The structural value-added model assumes the data-generating process is driven by the firm's interior solution of the Leontief model.

Under the usual structural value-added model, the firm's marginal product of intermediate material is predictably constant over time. My extension generalises the structural value-added model by allowing the firm to predict changes in the marginal product of intermediate material over time. Next, it becomes apparent that the structural value-added model illustrates the previously described capacity narrative.

The firm's interior solution has $M_{it} = \mathcal{C}^{-1}\left(\exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\right)$ because of $\mathcal{C}$'s strict monotonicity. Here, the $\mathfrak{g}$ function is $\log\left(\mathcal{C}^{-1}\left(\exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\right)\right)$ as not dependent of $\alpha^0_{g_i^0 t}$. Furthermore, the interior solution also implies $Y_{it} = \exp\left(\epsilon_{it} + \alpha^0_{g_i^0 t} + \omega_{it}\right) F\left(K_{it}, L_{it}\right)$.[11] So the structural value-added model assumes the data generating process is based on firms applying the interior solution. When $\mathcal{C}$ is convex then it overcomes many concerns raised by Gandhi, Navarro, and Rivers (2017b) about the firm achieving the interior solution.

Investment - v_{it} as logged investment

Economic models frequently assume that capital is subjected to some adjustment cost or delay with the installation. Hence, the firm's investment decision h is not sensitive to short-term productivity changes. Then $\alpha^0_{g_i^0 t}$ can stand as short-term productivity fluctuations. Furthermore, ω_{it} can stand for more persistent productivity changes.

In this environment, the firm's capital input is not correlated with $\alpha^0_{g_i^0 t}$. However, when the firm's labour input faces no dynamic constraints; labour is correlated with $\alpha^0_{g_i^0 t}$. For concreteness, I consider an example of Cobb-Douglas technology and a price-taking firm with its period t's capital investment as only effective at the start of period $t+1$. With the predetermined capital K_t, the firm chooses labour L_{it} to maximize its profit $\Pi_t\left(K_{it}\right) =$

[11] Under the interior solution, the semiparametric regression is

$$y_{it} = \log\left(\mathcal{C}\left(M_{it}\right)\right) + \alpha^0_{g_i^0 t} + \epsilon_{it}.$$

Which makes it just a semiparametric model of just intermediate material M_{it}. However, by generalising the gross production function to

$$Y_{it} = \exp\left(\alpha^0_{g_i^0 t} + \epsilon_{it}\right) \min\{\mathcal{C}\left(M_{it}, K_{it}, L_{it}\right), \exp\left(\omega_{it}\right) F\left(K_{it}, L_{it}\right)\},$$

can provide the semiparametric regression as

$$y_{it} = \log\left(\mathcal{C}\left(M_{it}, K_{it}, L_{it}\right)\right) + \alpha^0_{g_i^0 t} + \epsilon_{it},$$

under the interior solution. Then the regression is a function of (M_{it}, K_{it}, L_{it}).

$p\mathbb{E}\left[\exp\left(\epsilon_{it}\right)\right]\exp\left(\alpha_{g_i^0 t}+\omega_{it}\right)\left(K_{it}^{\beta_k}L_{it}^{\beta_k}\right)-rK_{it}-wL_{it}$, where p and (w,r) are output price and factor prices, respectively.

The standard optimization yields logged labour as $l_{it}=c_{Lit}+\dfrac{\alpha^0_{g_i^0 t}}{1-\beta_l}+\dfrac{\beta_k}{1-\beta_l}k_{it}$, where $c_{Lit}=\dfrac{1}{1-\beta_l}\log\left(\dfrac{p\beta_l}{w}\mathbb{E}\left[\exp\left(\epsilon_{it}\right)\right]\exp\left(\omega_{it}\right)\right)$. Hence, conditional on k_{it} and ω_{it}, l_{it} is still mean-dependent of $\alpha^0_{g_i^0 t}$. After the labour choice, the firm invests in capital, $\exp\left(\upsilon_{it}\right)$, to maximize its discounted δ_1 future profit, $\sum_{T=t}^{\infty}\delta\mathbb{E}\left[\Pi_{t+1}\left(K_{it+1}\right)|\mathcal{I}_t\right]$ subjected to the capital accumulation dynamic, $K_{is+1}=\left(1-\delta_2\right)K_{is}+\exp\left(\upsilon_{is}\right)$, where $s\geqslant t$ and δ_2 is the depreciation rate. Suppose $\alpha^0_{g_i^0 t}$ is independently and identically distributed over time within the group. Then the future profit is constant over $\alpha^0_{g_i^0 t}$ and, in turn, the υ_{it} is constant of $\alpha^0_{g_i^0 t}$. Thus the investment function $\mathfrak{g}$ does not depend on $\alpha^0_{g_i^0 t}$. Finally, Olley and Pakes (1996) discusses how υ_{it} can be monotonic with respect to ω_{it} in the setup here.

Comparison Against the Alternatives

The fixed effects model is the first proposed solution to address this simultaneity problem. However, it requires the observed productivity to be time-invariant. Here, none of the firm's productivity components has to be time-invariant. Furthermore, the fixed effects estimator is known to produce unreasonably low capital coefficient estimates, as reviewed by Griliches and Mairesse (1998). The suspect is the fixed effects' within-transformation exacerbates attenuation bias from classical measurement error in the capital.

Griliches and Hausman (1986) shows attenuation bias increases as information is swept out of the regressors. The fixed effects estimator induces within-transformation, and the information loss is most severe when the regressors are highly serially correlated. As noted by Levinsohn and Petrin (2003), many firms make lumpy capital-investment decisions, and, as a consequence, capital is likely to be highly serially correlated. Fortunately, the grouped fixed effect estimator avoids within-transformation but applies between-transformation. Hence, the grouped fixed effects estimator is more resilient against attenuation bias, compared to fixed effects, when the between-firm variation is significantly larger than the within-firm variation. Section 5 re-visits this point and shows the between-firm variation is more pronounced in the Chilean data.

As an alternative to the proxy variable setup, the dynamic panel approach avoids the inversion setup, but it assumes the firm treats ω_{it} as an autoregressive process. Furthermore, it estimates the autoregressive process with moment conditions. In summary, the dynamic panel method avoids the proxy variable assumptions for a simple autoregressive ω_{it} and using more moment conditions. More recently, Cheng, Schorfheide, and Shao (2019) shows how to estimate the dynamic panel approach with heterogeneous productivity means at the group level. In contrast to the fixed effects model, the group specification does not suffer the incidental parameter bias problem.

The other traditional avenues are to use either the firm's first-order condition behaviour or input prices as instruments. Imposing the firm's first-order condition either requires assuming perfect competition or the knowledge of each firm's output demand curve. Assuming perfect

competition is not appropriate in applications where firms have market power, as in De Loecker, Eeckhout, and Unger (2018), De Loecker and Scott (2017), and De Loecker and Warzynski (2012). Recovering the firm's output demand curve requires additional consumer preference assumptions and the demand side's data set. For the input prices instrument approach, the firm's specific input prices must be available and provide valid exogenous variation. As both Ackerberg, Caves, and Frazer (2015) and Gandhi, Navarro, and Rivers (2017a) notes, having valid and reliable instrumental input prices is not typical in the data. In summary, these alternatives place a much higher demand for what is available in the data.

The proxy variable's niche is the combination of allowing a general Markov process, being a minimalist in both data requirement and making assumptions on the market structure, and utilising cross-sectional variation to control for ω_{it}. My extension introduces firms' correlated productivity while keeping many of the proxy variable's advantages.

Conditional Method of Moments

To cover the general production function problem, I set up a conditional method of moments problem. The interest is to estimate the parameter τ ($\in \mathcal{T}$) and its observable variables are generically denoted as w_{it} ($\in \mathfrak{W} \subset \mathbb{R}^{d_5}$) - potentially to include y_{it}, x_{it} or z_{it} and their lagged values.

Define $\underline{\omega}(z_{it}, \tau) := m(z_{it}) - f(z_{it}, \tilde{\tau}) - \nu$. Then the firm's Markov prediction can be expressed as,

$$R(\underline{\omega}(z_{it}, \tau), \tau) = \mathbb{E}\left[\underline{\omega}(z_{it+1}, \tau) \mid \underline{\omega}(z_{it}, \tau)\right]. \tag{26}$$

Since z_{it} is stationary from the partially linear theory's assumption, the function R is not a function of t. R is the first order autoregression of $\underline{\omega}(z_{it}, \tau)$ and the estimation procedure applies basis approximation to estimate R. Without loss of genearality, the R function can be treated as a function of (w_{it}, τ). Then, τ^0 solves the set of moment conditions,

$$\mathbb{E}\left[\mathfrak{m}\left(w_{it}, \theta^0, m(z_{it}), \alpha^0_{g_i^0 t}, \tau, R(w_{it}, \tau)\right)\right] = 0. \tag{27}$$

This setup covers moments built from higher ordered lagged inputs. The criterion function is denoted as,

$$\hat{\mathfrak{M}}_{NT}(\tau)$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\mathfrak{m}\left(w_{it}, \hat{\theta}, \hat{m}(z_{it}), \hat{\alpha}_{\hat{g}_i t}, \tau, \hat{R}(w_{it}, \tau)\right)\right)' W \left(\frac{1}{T}\sum_{t=1}^{T}\mathfrak{m}\left(w_{it}, \hat{\theta}, \hat{m}(z_{it}), \hat{\alpha}_{\hat{g}_i t}, \tau, \hat{R}(w_{it}, \tau)\right)\right),$$

for some non-stochastic[12] positive definite weight matrix W. Then the second-step estimator

$$\hat{\tau} \in \arg\min_{\tau \in \mathcal{T}} \hat{\mathfrak{M}}_{NT}(\tau). \tag{28}$$

In the next subsection, I provide sufficient conditions for $\hat{\tau}$ to be consistent. The strategy is to verify Chen, Linton, and Keilegom (2003)'s high-level assumptions for their Theorem 1 in my setup, where there are both time and cross-sectional dependence. Furthermore, I need to show $\hat{R}$ as uniformly consistent over w_{it} and τ. The problem is non-trivial because $\hat{R}$ is estimated by series where both its outcome and regressor depend on the parameter τ.

[12]Extending W to be stochastic is straightforward when W is asymptotically convergent in probability to a positive definite matrix. For simplicity, I omit this extension.

3.2 Asymptotic Theory

Sufficient conditions for $\hat{\tau}$

Here, I provide the sufficient conditions for consistency of the second-step estimator. The objective is to estimate τ in presence of the nuisance parameter $\mathfrak{h}(w_{it}, z_{it}, \tau) = (\theta, m(z_{it}), \alpha_{NT}(g), R(w_{it}, \tau))$, where $\alpha_{NT}(g) = \left\{\{\alpha_{g_i t}\}_{t=1}^{T}\right\}_{i=1}^{N}$. In this notation,

$$\alpha_{NT}^{0}\left(g^{0}\right) = \left\{\left\{\alpha_{g_i^0 t}^{0}\right\}_{t=1}^{T}\right\}_{i=1}^{N}, \hat{\alpha}_{NT}(\hat{g}) = \left\{\{\hat{\alpha}_{\hat{g}_i t}\}_{t=1}^{T}\right\}_{i=1}^{N},$$

$$\hat{\mathfrak{h}} = \left(\hat{\theta}, \hat{m}(z_{it}), \hat{\alpha}_{NT}(\hat{g}), \hat{R}(w_{it}, \tau)\right), \text{ and } \mathfrak{h}^{0}(w_{it}, z_{it}, \tau) = \left(\theta^{0}, m^{0}(z_{it}), \alpha_{NT}^{0}\left(g^{0}\right), R^{0}(w_{it}, \tau)\right).$$

Besides R, the nuisance parameters are inherited and estimated from the first step partially linear model. Here, I assume estimator of R as uniformly consistent but the next part provides the sufficient conditions for it to happen. Appendix-B contains the proof.

Assumption S 1. *(Identification)*

1. *For any $\delta > 0$, there exists an $\epsilon(\delta) > 0$ such that*

$$\inf_{\|\tau - \tau^{0}\| > \delta} \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \mathfrak{h}^{0}(w_{it}, z_{it}, \tau)\right)\right]' \mathbb{E}\left[\mathfrak{m}\left(w_{it}, \tau, \mathfrak{h}^{0}(w_{it}, z_{it}, \tau)\right)\right] > \epsilon(\delta),$$

   for any i and t.

In the model's setup, the true parameter τ^0 solves the conditional moments. Assumption S 1 ensures τ^0 does not suffer the weak identification issue with using the conditional moments.

Assumption S 2. *(Compactness)*

1. *$\mathcal{T}$ is compact and has a non-empty interior containing τ^0.*
2. *The supports $\mathfrak{W}$ and $\mathcal{Z}$ are compact.*

Assumption S 2.1 precludes the analysis in dealing with the boundary value problem. As in Chen, Linton, and Keilegom (2003), I require the sample analogs of the conditional moments to converge to its population criterion uniformly. The Assumption S 2.2 compactness assumption helps to ensure this convergence can happen in the N as comparably larger than the T paradigm. For the application, the compactness does not introduce new constraints. There is no additional random variable w, and the basis function is a polynomial. Generally speaking, polynomials can only achieve uniform approximation over compact sets.

Furthermore, the compact supports also help to turn the criterion into Lipschitz. The Lipschitz condition is a convenient assumption to achieve uniform convergence, as mentioned by Chen, Linton, and Keilegom (2003).

Assumption S 3. *(Moments and Dependency)*

1. *$\left\{\left(w_{it}, z_{it}, \alpha_{g_i^0 t}^{0}\right)\right\}_{t=1}^{\infty}$ has an alpha mixing coefficient $\rho_i^{w,z,\alpha}(t)$ satisfying $\sup_{1 \leqslant i \leqslant T} \rho_i^{w,z,\alpha}(t) < C^{w,z,\alpha} \exp(-p_1 t)$, for some constant $C^{w,z,\alpha}$ and $p_1 > 0$.*

2. *For each* i, $\left(w_{it}, z_{it}, \alpha^0_{g_i^0 t}\right)$ *is a stationary process over* t.

Assumption S 3.1 is a weak dependency for the joint distribution of $\left(w_{it}, z_{it}, \alpha^0_{g_i^0 t}\right)$. The partially linear model does not necessarily have w_{it}. Hence, the previous weak dependency conditions alone do not necessarily imply Assumption S 3.1. z_{it} as a stationary process is already covered by Assumption 9, but is re-stated in Assumption S 3.2 for the ease of reference in the proof.

For the production function setup, these weak dependency conditions hold if intermediate material (or investment), labour, and capital are functions of independent state variables satisfying these weak dependency conditions. The simplest example is when the firm has no dynamic constraints on input choices and faces prices that are mutually independent processes and weakly time dependent.

However, it is natural to assume that capital faces dynamic constraints, then capital is also a state variable but it has a natural autoregressive transition. Then capital can be weakly dependent if the firm's investment function is sufficiently weakly time dependent. Furthermore, all state variables are no longer mutually independent because of capital as an additional state variable. So checking the weak time dependency for labour and intermediate material become more involved than before. It is useful to come up with simple sufficiency conditions for future research.

The estimation problem feeds the conditional moments with the first-step estimators, rather than the actual parameters. The estimation error is measured by the following metric,

$$d\left(\mathfrak{h}, \mathfrak{h}'\right) = \left\|\theta - \theta'\right\| + \sup_{z \in \mathcal{Z}} \left|m\left(z\right) - m'\left(z\right)\right| + \sup_{1 \leqslant i \leqslant N; 1 \leqslant t \leqslant T} \left|\alpha'_{g_i' t} - \alpha_{g_i t}\right| + \sup_{\tau \in \mathcal{T}} \sup_{w \in \mathcal{W}} \left\|R\left(w, \tau\right) - R'\left(w, \tau\right)\right\|.$$

Assumption S 4. *(Regularity)*

1. R^0 *is continuously differentiable.*
2. m^0 *is continuous.*
3. $\mathfrak{m}$ *is continuously differentiable over* $\mathbb{R}^{3+d_2+d_3+d_4}$.

These regularity conditions and the previous compact support assumptions complete $\mathfrak{m}$ as Lipschitz. For the production function case, differentiability is easy to verify for m^0 and $\mathfrak{m}$. For example, m^0 is differentiable when the parametric production function is differentiable, and the firm's conditional demand of the proxy variable has a nowhere vanishing derivative.

Assumption S 5. *(Rate)*

1. $\dfrac{\log\left(N\right)}{T} \to 0$, *as* $N, T \to \infty$.
2. $d\left(\hat{\mathfrak{h}}, \mathfrak{h}^0\right) = o_p\left(1\right)$.

Assumption S 2.2 is immediate from the results in Theorem 4 and the presumption of having a consistent estimator for R^0. Assumption S 2.1 is compatible with the larger N than T setup as the log function is slowly varying.

Theorem S 1. *Under Assumption 2, S 1, S 2, S 3, S 4, and S 5,*

$$\hat{\tau} \xrightarrow{P} \tau^0,$$

as $N, T \to \infty$.

The case of a stochastic matrix W is not formally covered here. However, Theorem S 1's argument can be adapted to hold when the stochastic W satisfies the high-level conditions specified in Chen, Linton, and Keilegom (2003)'s Corollary 1.

The paper currently does not provide the theory to make inference on $\hat{\tau}$. Verifying the sufficient conditions leading up to Chen, Linton, and Keilegom (2003)'s Theorem 2 would provide a central limit theorem result. One important condition is to have $\hat{\mathfrak{h}}$ to converge at the $(NT)^{\frac{1}{4}}$ rate. The partially linear theory shows that this can happen for $\hat{m}$ and $\hat{\theta}$. Furthermore, the next subsection provides sufficient conditions for $\hat{R}$ to do so. However, $\hat{\alpha}_{\hat{g}_i t}$'s rate is unknown and to proceed forward may require dropping moment conditions using $\hat{\alpha}_{\hat{g}_i t}$. Furthermore, Chen, Linton, and Keilegom (2003) also requires a Donsker condition on the criterion function, and they only provide a reference to verify this condition for cross-sectional data with independence. However, section 4 shows that the bootstrap confidence interval provides the correct coverage in simulation when T is large. It appears the normality approximation and bootstrap standard errors can be used for inference, even under the cross-sectional dependence from α^0 and the time-series dependence for each unit.

Sufficient conditions for $\hat{R}$

Here, I provide sufficient conditions to verify the uniform consistency of $\hat{R}$ - assumed by the Conditional Method of Moments. As mentioned previously, I consider $\hat{R}$ as a non-parametric estimator using basis $\{b^L\}_{L=1}^{\infty}$.

The proof uses a similar argument presented in Ai and Chen (2003) Lemma 1's proof. However, their setup is large N asymptotics, and only their dependent is a function of the unknown parameter. Some of the presented assumptions here overlaps the ones in the section 2. However, there is no issue of conflict and repeating them helps for reference purposes. Appendix-C contains the proof.

Assumption R 1. *(Dependency) Let C^f be some positive constant.*

1. *Conditional on α, $\{z_{it}\}_{t=1}^{\infty}$ is independent over i. Furthermore, it is unconditionally stationary over t.*

2. *Conditional on α and for each i, the process z_{it} has an alpha mixing coefficient $\rho_i^z(\alpha, t)$. Furthermore, $\sup_{i \in \{1,\ldots,N\}} \sum_{t=0}^{\infty} \left(\rho_i^z(\alpha, t)\right)^{\frac{1}{3}} < C^f$.*

Assumption R 1 is already covered by Assumption 9 in the partially linear model.

Define $\underline{\omega}(z_{it}, \tau) := m(z_{it}) - f(z_{it}, \tilde{\tau}) - \nu$. By Assumption R 1.1, $\underline{\omega}(z_{it}, \tau)$ is independent over i, conditional on α. Note that $R(\cdot, \tau)$ is the regression of $\underline{\omega}(z_{it}, \tau)$ against its first lag, i.e.

$$R(\underline{\omega}(z_{it}, \tau), \tau) = \mathbb{E}\left[\underline{\omega}(z_{it+1}, \tau) \mid \underline{\omega}(z_{it}, \tau)\right].$$

The R function's second argument, τ, captures the fact of $\underline{\omega}(z_{it}, \tau)$'s distribution varying with τ. The residual $v_{it}^{\tau} := \underline{\omega}(z_{it}, \tau) - R(z_{it}, \tau)$ is mean zero by construction. Though $v_{it}^{\tau^0}$ is serially uncorrelated in production theory, misspecified τ $(\neq \tau^0)$ induces serial correlation of v_{it}^{τ}.

R function can be re-parameterized as a function of just (z_{it}, τ). This version best fits the R's description in the GMM step with $w_{it} = z_{it}$. However, for asymptotic analysis, it is more natural to treat R as function of (w, τ) due to using the generated regressor $\hat{m}$.

Assumption R 2. *(Compactness) $\mathcal{Z}$ and $\mathcal{T}$ are compact.*

This assumption is already covered by Assumption S 2.

Assumption R 3. *(Smoothness)*

1. *R is continuously differentiable in (ω, τ).*
2. *f is continuously differentiable.*
3. *m is continuously differentiable.*

Assumption R 3's only addition over Assumption S 4 is the production function is also continuously differentiable. Assumption R 3 turns $\underline{\omega}(z_{it}, \tau)$ as continuously differentiable. In conjunction with Assumption R 2, $\underline{\omega}(\mathcal{Z} \times \mathcal{T})$ is compact in $\mathbb{R}$. Thus $\underline{\omega}(\mathcal{Z} \times \mathcal{T})$ is contained in the interior of a larger compact set $\mathcal{W}$, with $B(\underline{\omega}(z, \tau), \delta_W) \subset \mathcal{W}$ for a constant δ_W and any $(z, \tau) \in \mathcal{Z} \times \mathcal{T}$. This uniform radius helps to bound the generated regressor's error in the proof.

Assumption R 4. *(Approximation)*

1. *There exists a sequence of functions $\{r^{0,L}\}_{L=1}^{\infty}$ and $\mu_R > 0$ such that*

$$\sup_{\omega \in \mathcal{W}} \sup_{\tau \in \mathcal{T}} \left| b^L(\omega)' r^{0,L}(\tau) - R(\omega, \tau) \right| = O\left(L^{-\mu_R}\right).$$

2. *b^L is continuously differentiable on $\mathcal{W}$.*

Assumption R 4.1 can be readily verified for power series. R's smoothness assumption allows the (w, τ) power series to uniformly approximate R over $\mathcal{W} \times \mathcal{T}$. Then $r^{0,L}(\tau)$ can be constructed by combining the approximating coefficients with the factors involving orders of τ. Assumption R 4.2 implies b^L and $\dfrac{db^L}{d\omega}$ are bounded on $\mathcal{W}$. So there exists a sequence of monotonic bounds, $\left\{\xi_L^b\right\}_{L=1}^{\infty}$, satisfying $\sup_{l \in \{1,...,L\}} \max\left\{\sup_{\omega \in \mathcal{W}} \left|b_l^L(\omega)\right|, \sup_{\omega \in \mathcal{W}} \left|\dfrac{db_l^L}{d\omega}(\omega)\right|\right\} < \xi_L^b$. Then they imply $\max\left\{\sup_{\omega \in \mathcal{W}} \left\|b^L(\omega)\right\|, \sup_{\omega \in \mathcal{W}} \left|\dfrac{db_l^L}{d\omega}(\omega)\right|\right\} < \sqrt{L}\xi_L^b$. Assumption R 4 also provides the smoothness conditions for the R's version as function of (z_{it}, τ).

To ensure the rank condition, the standard setup places eigenvalue restrictions on the matrix $\mathbb{E}\left[b^L(\underline{\omega}(z_{it}, \tau)) b^L(\underline{\omega}(z_{it}, \tau))'\right]$. Due to cross-sectional dependence, here analogous assumption is to place the similar restrictions on the same expectation but conditioning on α. But, conditioning α, $\underline{\omega}(z_{it}, \tau)$ is not stationary but time-dependent.

Assumption R 5. *(Rank Condition) There exist a constant C and a sequence of positive definite matrix of functions $\left\{\psi^{bb,L}(\alpha, \tau, T)\right\}_{L=1}^{\infty}$ such that*

1. $\left\|\psi^{bb,L}(\alpha, \tau, T) - \dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{E}\left[b^L(\underline{\omega}(z_{it}, \tau)) b^L(\underline{\omega}(z_{it}, \tau))' | \alpha\right]}{NT}\right\| < \dfrac{C}{\sqrt{N}}.$

2. $\dfrac{\sum_{l=1}^{L} \mathbb{E}\left[\frac{1}{\lambda_{lL}(\alpha, \tau, T)}\right]}{(NT)^{r^\psi}} < C$ *for some* $r^\psi \in [0, \dfrac{1}{4})$, *where* $\lambda_{1L}(\alpha, \tau, T), ..., \lambda_{LL}(\alpha, \tau, T)$ *are the matrix* $\psi^{bb,L}(\alpha, \tau, T)$*'s eigenvalues.*

Assumption R 5 places the eigenvalue restrictions on the matrix $\psi^{bb,L}(\alpha,\tau,T)$, which has to be time-varying and dependent on α. The matrix $\psi_L^{bb}(\alpha,\tau,T)$ is just

$$\sum_{g=1}^{G^0}\kappa_g\frac{\sum_{t=1}^{T}\mathbb{E}\left[b^L\left(\underline{\omega}\left(z_{it},\tau\right)\right)b^L\left(\underline{\omega}\left(z_{it},\tau\right)\right)'\middle|\alpha,g_i^0=g\right]}{T},$$

when the moments are identical within the group.

Assumption R 5.2 uniformly (over τ) controls the relative size of $\left[\psi^{bb,L}(\alpha,\tau,T)\right]^{-1}$'s norm to the sample size. The issue is to ensure the convergence rate as not dependent on τ. This restriction ensures sufficient data signal is available for all $\tau\in\mathcal{T}$. Typically, $\psi^{bb,L}$'s larger eigenvalues increase with L and is a measure of the signal. When there is sufficient signal (i.e., the sum of the expected inverse eigenvalues is uniformly bounded); Assumption R 5.2 is satisfied with $r^{\psi}=0$. When group members have identical moments, λ_{lL} does not change with N. Then Assumption R 5.2 is satisfied; when L increases cautiously relative to the size of N.

Assumption R 6. *(Rates) Let* $\mathbb{E}\left[\sup_{z\in\mathcal{Z}}|\widehat{m}(z)-m(z)|\right]=\Delta$, $\mathfrak{d}_{NTL}=\log\left(L\left(\xi_L^b\right)^2(NT)^{r^{\psi}}\right)$, *and* $\mathfrak{v}_{NTL}=d_3\log\left(L\left(\xi_L^b\right)^2(NT)^{\frac{1}{4}+r^{\psi}}\right)$. *As* $N,T,L\to\infty$,

1. $\frac{T}{N}\to 0$.
2. $\frac{N^{\frac{1}{4}+r^{\psi}}L\left(\xi_L^b\right)^2}{L^{\mu_R}}\to 0$.
3. $\max\left\{\frac{\mathfrak{v}_{NTL}\left(\xi_L^b\right)^4L}{(NT)^{\frac{1}{2}-2r^{\psi}}},\frac{\mathfrak{d}_{NTL}\left(\xi_L^b\right)^4}{N^{1-r^{\psi}}},\frac{\mathfrak{v}_{NTL}\left(\xi_L^b\right)^2\sqrt{L}}{(N)^{\frac{1}{2}-2r^{\psi}}}\right\}\to 0$.
4. $(NT)^{\frac{1}{4}}\xi_L^b\Delta\to 0$.
5. $L\left(\xi_L^b\right)^2\Delta\to 0$.

Assumption R 6 provides the conditions for the series estimator to uniform converge at the $(NT)^{\frac{1}{4}}$ rate. For just uniform consistency, the proofs can be adapted by using weaker versions of Assumption R 6.2, R 6.4, and R 6.5 by scaling the rates with $\frac{1}{(NT)^{\frac{1}{4}}}$. Furthermore, Assumption R 6.3 can weakened to $\frac{\mathfrak{v}_{NTL}\left(\xi_L^b\right)^4L}{(NT)^{1-2r^{\psi}}}\to 0$, $\frac{\mathfrak{v}_{NTL}\left(\xi_L^b\right)^2(T)^{\frac{1}{4}+r^{\psi}}\sqrt{L}}{(N)^{1-r^{\psi}}}\to 0$, and $\frac{\mathfrak{d}_{NTL}\left(\xi_L^b\right)^4}{N^{1-r^{\psi}}}\to 0$.

Assumption R 6.1's large N over T setup ensures Assumption R 5.1's cross-sectional heterogeneity bound is asymptotically negligible. Assumption R 6.3 and R 6.4 ensure the series estimator uniformly converges at $(NT)^{\frac{1}{4}}$ rate in absence of $\hat{m}$'s estimation error. Assumption R 6.4, R 6.5, and R 6.6 requires the rate of $\hat{m}$'s estimation error to dissipate. With Theorem 4, $\Delta=O_p\left(\xi_K K^{-\mu}+\xi_K\sqrt{K}\Pi_K N^{-1}\right)+O_p\left(\xi_K^2\frac{\sqrt{K}}{\sqrt{NT}}\right)$. The term $O_p\left(\xi_K\sqrt{K}\Pi_K N^{-1}\right)$ comes from the cross-sectional heterogeneity in moments.

Theorem R 1. *Under Assumption R 1, R 2, R 3, R 4, R 5, and R 6,*

$$\sup_{\tau\in\mathcal{T}}\sup_{\omega\in\mathcal{W}}(NT)^{\frac{1}{4}}\left|\hat{R}(\omega,\tau)-R(\omega,\tau)\right|=o_p(1).$$

4 Monte Carlo

The Monte Carlo section covers two sets of simulation exercises. The first set simulates the coverage probability of $\hat{\theta}$'s confidence interval in a partially linear dynamic panel setting. The confidence intervals use the covariance formula and the asymptotic normality result presented in the asymptotic theory section. Then the second set studies the two-step estimator in the production function setting. For the first step, the simulation benchmarks the information criterion's performance and provides comparative statics in studying classification error. For applied interest, I also simulate the bootstrap confidence interval's coverage for the two-step estimator.

Generally, the classification error vanishes, and the coverage probabilities is close to the nominal values as the number of periods increases. Similarly, the information criterion overwhelmingly selects the correct model. Each simulated trial computes the estimator by using four hundred different group initialisations. The unknown function is approximated by a cross-validated polynomial, unless specified otherwise. Appendix D describes the cross-validation procedure. Furthermore, the minimum and maximum considered orders are the third and seventh degrees, respectively.

4.1 Partially Linear: Dynamic Panel

The dynamic panel model is

$$y_{it} = \theta y_{it-1} + \phi\left(z_{it}\right) + \alpha^0_{g^0_i t} + \epsilon_{it} : \theta = 0.5, \tag{29}$$

where ϕ is a standard normal probability density function. This setup is a toy model of income growth, y_{it}, dependent of the level of inequality, z_{it}, and institutional effects, $\alpha^0_{g^0_i t}$. By modeling z_{it}'s effect through the normal density, the model implies either the lack of or excessive inequality is not desirable for growth. Intuitively, lack of inequality may stifle incentives to produce, and excessive inequality can impede innovative entrant firms to access resources.

There are four groups, and each has a stationary process α_{gt} with a unique mean as either 0, 0.25, 0.5, or 1. z_{it} is the sum of $\alpha_{g^0_i t}$ and another autoregressive process with zero mean. Furthermore, the model has heteroskedasticity $\epsilon_{it} \sim N\left(0, \min\{1, y^2_{it-1}\}\right)$, independently. Conditional on α^0, all processes are independent over the cross-section. Moreover, all first-order autoregressive processes are generated by standard normal innovations and have 0.7 as its autocorrelation coefficient. The data generating processes are initialised at the stationary values.

Each group has the same number of memberships, and ϕ is approximated by a cross-validated polynomial of z_{it}. The other parameters are estimated as described previously.

N_g	$T = 5$	$T = 10$	$T = 15$	$T = 20$
40	90.1% (12.35%)	94.6% (1.81%)	96.8% (0.27%)	95.4% (0%)
100	85.8% (19.43%)	96.7% (4.55%)	96.4% (1.03%)	96.8% (0.27%)
200	80.5% (19.82%)	96.1% (4.86%)	96.6% (1.17%)	96.5% (0.31%)

Table 1: Coverage Probability for 95% Nominal Confidence Interval for $\hat{\theta}$

The simulated results are tabulated from one thousand trials. The parenthesis reports the average classification errors[13] up to the second decimal, and the simulation shows the coverage is

[13]The estimated groups are only identified up to a permutation. To quantify the classification error, I match the estimated group with its members' modal true group g .

N_g	$T=5$	$T=10$	$T=15$	$T=20$
40	83.9% (17.35%)	90.2% (3.52%)	92.3% (0.77%)	91.3% (0.12%)
100	79.9% (19.43%)	91.8% (4.55%)	91.5% (1.03%)	91.7% (0.27%)
200	73.8% (19.82%)	91.2% (4.86%)	92.1% (1.17%)	92.2% (0.31%)

Table 2: Coverage Probability for 90% Nominal Confidence Interval for $\hat{\theta}$

close to the nominal value as the number of periods increase. Furthermore, the classification error also drops with the number of periods, just as the asymptotic theory predicts.

The reader may notice two patterns from the tables. First, the classification error tends to be higher with a larger N_g. A larger sample is more likely to populate the empirical distribution's tail. Furthermore, the tail observations are likely to be misclassified. This phenomenon explains the little increase of classification error with larger N_g.

Second, the coverage probability tends to be larger than the nominal value with larger sample size. Under more numerous observations, z_{it} has more variations to reveal the finite polynomial's approximation error on ϕ.

N_g	$K=1$	$K=3$	$K=5$	$K=7$
200	84.8%	94.9%	92.7%	91%

Table 3: Coverage Probability for 90% Nominal Confidence Interval for $\hat{\theta}$ when $T=30$

N_g	$K=1$	$K=3$	$K=5$	$K=7$
200	92.3%	97.9%	96.2%	96%

Table 4: Coverage Probability for 95% Nominal Confidence Interval for $\hat{\theta}$ when $T=30$

It appears that the approximation error is causing the upward distortion of the coverage probability. The tables show coverage moves towards the nominal value by increasing K. Hence, the simulation is consistent with the theory's asymptotic prediction.

4.2 Two-Step: Production Function

Here, I assess the finite sample performance of my production function estimator when the intermediate material acts as the proxy variable. Also, the Monte Carlo verifies the asymptotic of my information criterion and classification consistency results when a polynomial non-parametrically estimates m. The results serve both the interest of my production function application and the general use of my two-steps estimator.

My data-generating process is an extension of Ackerberg, Caves, and Frazer (2015)'s DGP1 used in their Monte Carlo. Their setup provides a simple solution to the firm's dynamic profit maximization problem and, then, simulates the data from firms' policy functions.

The structural value-added production function is Cobb-Douglas and the gross production function is

$$Y_{it} = \exp\left(\epsilon_{it} + \alpha^0_{g_i^0 t}\right) \min\left\{\mathcal{C}\left(M_{it}\right), \exp\left(\omega_{it}\right) K_{it}^{\beta_k} L_{it}^{\beta_l}\right\}, \tag{30}$$

where $\mathcal{C}(M_{it}) = M_{it} + M_{it}^2$. The objective is to estimate the output elasticity (β_k, β_l). $\mathcal{C}(M_{it})$'s monotonicity and convexity ensures the firm's interior solution. Furthermore, I can easily solve the optimal choice of M_{it} from $\mathfrak{F}(M_{it})$'s second-order polynomial form. Ackerberg, Caves, and Frazer (2015) used $\mathfrak{F}(M_{it})$ as a linear function of M_{it} but, after log-linearization, the filtering step's semiparametric form is exactly linear in logged M_{it}. From a non-parametric perspective, it is uninteresting to approximate logged M_{it} with a polynomial of itself. However, the group productivity extension works perfectly fine with the linear specification.

Here, I outline the productivity process, and Appendix D provides the full firm's decision problem, solves the policy functions, and other parametric details. There are three true groups, i.e., $G^0 = 3$. All three processes ϵ_{it}, $\alpha^0_{g_i^0 t}$, and η_{it} are mutually independent. Both ϵ_{it} and η_{it} are zero-mean independent processes over i but only ϵ is independent over t. η_{it} is a stationary first-order autoregressive process. The simulation generates firms' input choices based on the solved policy functions.

After applying the log transformation to the firm's interior solution of intermediate material,

$$y_{it} = \log\left(e^{\upsilon_{it}} + e^{2\upsilon_{it}}\right) + \alpha^0_{g_i^0 t} + \epsilon_{it}, \tag{31}$$

where $\upsilon_{it} = \log(M_{it})$. So I use a cross-validated polynomial of υ_{it} to non-parametrically estimate $\log\left(e^{\upsilon_{it}} + e^{2\upsilon_{it}}\right)$. For simplicity, I estimate R parametrically in the second stage.

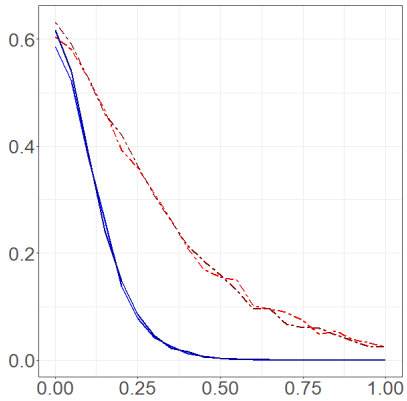
In the absence of $\log\left(e^{\upsilon_{it}} + e^{2\upsilon_{it}}\right)$, the classification problem is on $\alpha^0_{g_i^0 t} + \epsilon_{it}$. So intuitively, the classification of g_i^0 is an easier problem when $\alpha^0_{g_i^0 t} + \epsilon_{it}$ is more similar within the group than between groups. In my Monte Carlo setup, the classification precision is roughly positively associated with the ratio

$$\min_{g,g':g\neq g'} \frac{2\left(\sigma^2_{\alpha_g} - \sigma_{\alpha_g,\alpha_{g'}}\right) + \left(\mu_g - \mu_{g'}\right)^2}{\sigma^2_\epsilon}, \tag{32}$$

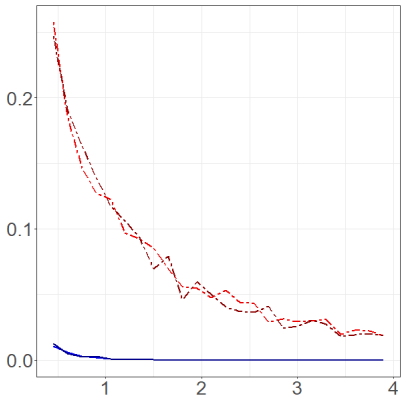
where $\mu_g, \sigma^2_{\alpha_g}$ and $\sigma_{\alpha_g,\alpha_{g'}}$, σ^2_ϵ are α_{gt}'s mean, variance, covariance with $\alpha_{g'}$, and ϵ_{it}'s variance, respectively. Appendix D provides a heuristic argument to why the ratio is informative in the partially linear semiparametric model.

The ratio suggests classification error decreases when different groups' α_{gt} become more dissimilar in mean or correlation. For my Monte Carlo Design 1, I model $\alpha^0_{gt} = w\alpha^*_{gt} + (1-w)\alpha^*_t$ - a convex combination of two mutually independent first-autoregressive processes, α^*_{gt} (group specific with $\mu_g \in \{-0.33, 0, 0.33\}$) and α^*_t (zero mean common trend). Classification error should fall as the grouped gross productivity processes become less correlated or more different in their means, i.e. when $w \to 1$.

The ratio also suggests the classification error decreases when $\dfrac{\sigma^2_{\alpha_g}}{\sigma^2_\epsilon}$ increases. That is to say, the classification precision improves when the firm's surprise productivity change ϵ_{it} comparably lowers in uncertainty. More predictable production environment should yield better classification estimates. For my Monte Carlo Design 2, I set $w = 1$ and vary $\dfrac{\sigma^2_{\alpha_g}}{\sigma^2_\epsilon}$ by increasing $\sigma^2_{\alpha_g}$.



Design 1



Design 2

Figure 1: Y-axis: Average Classification Error | X-axis: w and $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon}$ for Design 1 and 2, respectively.

Solid/Blue line: $T = 20$. Dashed/Red line: $T = 5$. Blue/Red: $N_g = 100$ (Number of observations for each group). Dark Blue/Dark Red: $N_g = 300$.

The Figure 1 estimates the average classification error at different parametric values, based on four hundred trials. And the simulated curves verify the previous predictions. Furthermore, the $T = 20$ curves are strictly lower than the $T = 5$ curves. Hence, they verify the asymptotic classification consistency result in the semiparametric model. Appendix D plots the difference between the mean-squared error of the elasticity estimates based on the true groups vs. the estimated groups. Their difference converge towards zero as the classification error vanishes. For Design 1, their difference also vanishes as $w \to 0$. Closer to $w = 0$, while classification identification is weaker, the elasticity estimates' bias is also smaller.

In the absence of classification error, production function literature has numerous studies of the two-step estimator's performance in Monte Carlo simulations. For brevity, I do not provide additional analysis of output elasticity estimates' mean-squared error, but next bootstrap results partially capture the estimator's performance in mean-squared error. For the inference of output elasticity estimates, I assess the bootstrap confidence interval's coverage. I do block bootstrap with each unit's time-series constituting a block. Each bootstrap sample constructs new output elasticity estimates at the second-step conditioning on the original sample's first-step estimates. Then I construct the bootstrap confidence interval from the normal critical values and the bootstrap empirical distribution's[14] standard errors.

T	N_g	β_k^0	β_l^0
5	100	81.5%	64%
5	300	73.5%	48%
20	100	96.5%	93.5%
20	300	95%	96%

Table 5: Coverage for the 95% Bootstrap Confidence Interval.

[14]Five hundred bootstrap samples construct the empirical distribution.

The table reports the coverage probability over four hundred trials and under Design 1 with $w = 0.5$. The coverage converges to the nominal value as the number of periods increases. From the classification perspective, this outcome is not surprising as classification error is near zero at $T = 20$. However, the bootstrap confidence interval is able to provide the near correct coverage despite the data's serial correlation and cross-sectional dependence. As already mentioned, the bootstrap theory to account for both serial correlation and cross-sectional dependence is not provided here. But the simulation provides an applied justification to use bootstrap standard errors for section 5.

For the information criterion, I set the penalty as $\frac{\lambda}{T^{\frac{1}{5}}}\hat{Q}_{G_{\max}}$[15]. Then λ is chosen by the following data-driven approach:

$$\hat{G} \in \underset{G \in \{1,...,G_{\max}\}}{\arg\min} \; IC_{\lambda^*}(G), \tag{33}$$

where $\lambda^* \in \underset{\lambda \in \mathcal{K}}{\arg\min} \left[\underset{G \in \{1,...,G_{\max}\}}{\min} IC_\lambda(G) \right]$ and $\mathcal{K} := \{0.18, 0.2, ..., 1.8, 2\}$. The asymptotic theory covers criterion's selection consistency over every $\lambda \in \mathcal{K}$ because $\mathcal{K}$ is a finite and fixed set. So the asymptotic result easily extends to the information criterion using the data-driven choice of λ.

At $w = 0.5$ for Design 1 or $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon} = 1$ for Design 2, the simulated classification error is around 17% when $T = 5$. With each group having 100 members, I assess the above information criterion's performance in the simulation at $w = 0.5$ for Design 1 and at $\frac{\sigma_{\alpha_g}}{\sigma_\epsilon} = 1$ for Design 2. Here, $G^0 = 3$ and the $G_{\max} = 6$. All group specifications use the same polynomial order, and the order is chosen by cross-validation based on the over-specification, $G = 6$. How to optimally and jointly determine the polynomial order and the true group G^0 is an avenue for future research.

Design	T	$\hat{G} = 1$	$\hat{G} = 2$	$\hat{G} = 3$	$\hat{G} = 4$	$\hat{G} = 5$	$\hat{G} = 6$
1	$T = 5$	0%	2.9%	84.7%	12.4%	0%	0%
1	$T = 20$	0%	0%	99.5%	0.5%	0%	0%
2	$T = 5$	0%	0.5%	86.5%	13%	0%	0%
2	$T = 20$	0%	0.1%	99.4%	0.5%	0%	0%

Table 6: Simulated frequency of $\hat{G}$'s realisation based on four hundred simulations at each specification.

The table shows the information criterion performs well in finite sample even under classification error. Furthermore, the table verifies the asymptotic consistency result of the information criterion. The error of both under-selection and over-selection decreases as T increases. As this information criterion performs well here, I use it for my empirical application.

5 Empirical Application

In this section, I illustrate the empirical performance of my production function estimator. The data set consists of Chilean manufacturing plants from 1987 to 1996 and is sourced from the census of Chilean manufacturing plants by Chile's Instituto Nacional de Estadística. It covers all firms with more than ten employees. My construction of capital, labor, and intermediate material follows

[15] $\hat{Q}_{G_{\max}}$ is the least-squared criterion evaluated at the parameters estimated from the $G_{\max}$ specification. Having the penalty scaled by $\hat{Q}_{G_{\max}}$ ensures the selection is invariant to the data's scale.

Gandhi, Navarro, and Rivers (2017a, 2017b). For studying production function estimation, this data set series has also been used by Levinsohn and Petrin (2003) and Lee, Stoyanov, and Zubanov (2019).

The data set is available from 1979 to 1996. However, Levinsohn and Petrin (2003) raises potential structural break concerns for the earlier years. I use the data from 1987 to avoid addressing those structural breaks. The four sectors Food Product (331), Wood Products (331), Textile (321), and Fabricated Metal Products (381) are within the data set's top five largest sectors and included in all the mentioned previous studies. I restrict my analysis to these four sectors.

Chile experienced significant economic growth from 1987 to 1996, and the years fall into the well-known Miracle of Chile period. The growth spurt occurred after significant economic reforms were implemented and can be interpreted as the economy's convergence to a new steady state. Here, I use my production function estimator to measure firms' productivity changes and distribution for the four sectors. In contrast to previous studies, I allow cross-correlation in firms' productivity and firms to have heterogeneous transition dynamics in productivity.

Within a sector, I assume all firms have the same output elasticity (β_k, β_l) and follow the Cobb-Douglas (structural value-added) production function,

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + \alpha^0_{g^0_0 t} + \epsilon_{it}, \tag{34}$$

with heterogeneous firm productivity. Using my production function estimator, I estimate the output elasticity for every four sectors. Here, the proxy variable is the firm's intermediate material choice. I use a second order polynomial of (k_{it}, l_{it}, v_{it}) for the filtering step and a third order polynomial of $\hat{\eta}_{it-1}$ to approximate its first-order Markov process.[16] Using the second order at the filtering step is common in the literature because it is parsimonious and has the translog production function interpretation. As shown in section 4, I use the information criterion to select the number of groups for each sector - the set of alternatives includes up to ten groups, and the polynomial is the second order. The main estimates use four groups for Food, and five groups for Metal and Textile, and six groups for Wood.

Selection:
Both Olley and Pakes (1996) and Griliches and Mairesse (1998) argue for using the unbalanced panel to mitigate the selection issue from the firm's entry and exit decisions. Beyond using the unbalanced panel[17], I do not address the selection issue. It is possible to include a near-verbatim Olley-Pakes style selection correction at the second GMM step, but that is beyond the paper's scope.

Sector	N	Median T_i	Mean T_i
Wood	236	4	5.26
Textile	320	6	6.41
Food	1140	8	6.79
Metal	436	5	5.90

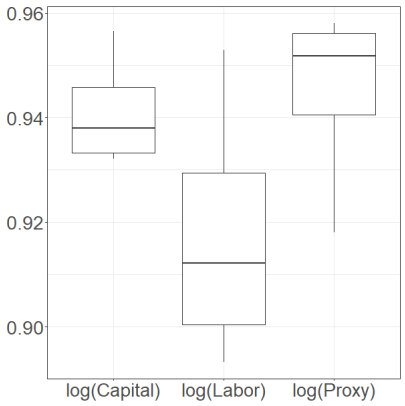
Table 7: *T_i is the ith firm's number of periods.*

[16]For robustness, all local optimization steps are done with over five hundred randomly selected initialization points.

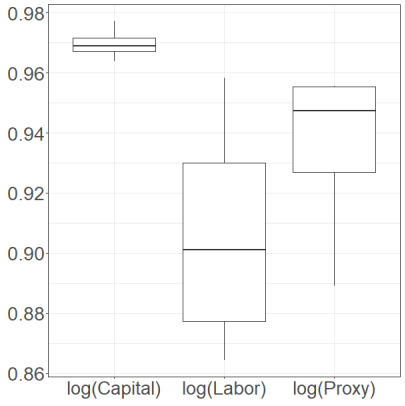
[17]It is not apparent on which T to substitute into the information criterion in this unbalanced panel setting. For simplicity, I use the firm's median number of periods.

All sectors have a sizable N dimension comparably to their T dimension. In section 4, precise classification can be achieved even when T is small but N is large. More specifically, this happens when the firm's productivity uncertainty is low, i.e. ϵ_{it} is less variable than α_{gt}. Or, when the different mean-level of α_{gt} are well-separated.

Measurement Error: Within-Variation vs Between-Variation



R^2 of the logged input's AR(1) model.



$$\frac{\text{Logged Input's Variation Between Firms}}{\text{Logged Input's Total Variation}}$$

Figure 2: Sources of Inputs' Variation.

The Figure 2 shows that the inputs are highly serially correlated, and the firms' between-variation of inputs dominates the firms' within-variation of inputs. As discussed in section 3, Griliches and Hausman (1986)'s intuition suggests the attenuation bias is less severe by applying between-transformation as opposed to within-transformation in the scenario here. On the measurement error issue, I note that the grouped fixed effects estimator should be more resilient to the attenuation bias's effect, as compared to the fixed effects estimator.

Evidence of Heterogeneous Productivity Groups

I find the model specification's fit is improved by including heterogeneous productivity groups. The evidence lies with the estimates of ϵ_{it}.

The productivity ϵ_{it} is unaccounted by the firm's input decisions because it is unpredictable during the firm's decision making. From the filtering step, the estimate $\hat{\epsilon}_{it}$ is invariant for all smooth Hicksian neutral technology choice - only the second GMM step imposes the production function's parametric form. Furthermore, the filtering step's estimates are robust to the standard selection concern - which arises in the second GMM step. Hence, the estimate $\hat{\epsilon}_{it}$ is reasonably robust and should be serially uncorrelated under the correct model specification.

The time-invariant Proxy Variable model is nested under the single group specification. The single group is misspecified, as shown in Figure 3 from the $\hat{\epsilon}_{it}$'s significant autocorrelation. Moreover, the autocorrelation faces an over 70% reduction by increasing the number of groups to four or five. The Figure 3 is consistent with the presence of predictable grouped productivity shock in the firm's decision environment. For completeness, I also estimate the time-varying Proxy Variable model with the same second-order polynomial. Even in there, $\hat{\epsilon}_{it}$'s AR(1) model has a R^2 of $0.625, 0.748, 0.631$, and 0.707 for Wood, Food, Metal, and Textile, respectively.[18]

[18]On the over-fitting case, the time-varying Proxy variable model is more parameterized than my het-

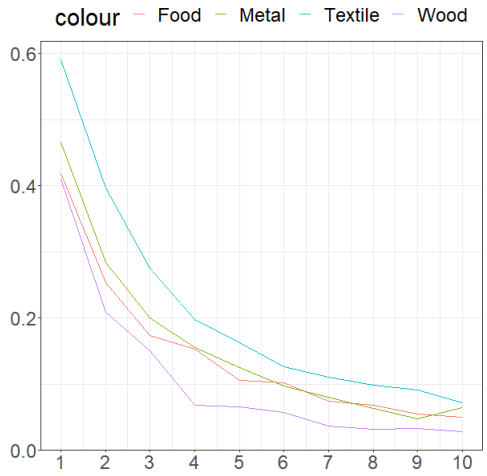


Figure 3: The R^2 of $\hat{\epsilon}_{it}$ AR(1) Model over G groups. Selected G is 4 for Food, 5 for Metal and Textile, and 6 for Wood.

The serial correlation of the proxy variable model's $\hat{\epsilon}_{it}$ has also been documented by Kasahara, Schrimpf, and Suzuki (2017), for the Japanese Machine Industry, and by Lee, Stoyanov, and Zubanov (2019), for Danish manufacturing firms.

As an alternative interpretation, Ackerberg, Caves, and Frazer (2015) models ϵ_{it} as serially correlated measurement errors on output. These measurement errors are innocuous only if they do not correlate with the measurement of inputs. However, I find the group with higher productivity tends to choose more capital input.[19] Capturing only innocuous measurement errors is not what drives down ϵ_{it}'s serial correlation in the graph. Nevertheless, the measurement error can be the cause behind the residual serial correlation after modeling the groups.

Groups' Composition:

Sector	G1	G2	G3	G4	G5	G6
Food	20.282	47.914	25.914	5.891	N/A	N/A
Metal	1.946	18.288	27.743	36.459	15.564	N/A
Textile	10.434	26.621	35.592	23.452	3.901	N/A
Wood	13.699	3.143	29.976	33.441	11.604	8.139

Table 8: Percentage of the sector's firm in each group. Groups are ordered in increasing mean level of $\hat{\alpha}_{gt}$.

erogeneous groups specification. My specification only adds one intercept per group for an additional year. However, the time-varying proxy variable model adds a new complete set of polynomial coefficients. My model is a parsimonious way to include heterogeneous productivity over time and different firms.

[19]Stacked barplots of groups' mean level inputs are available in Appendix E.

Sector	G1	G2	G3	G4	G5	G6
Food	2.179	13.619	56.136	28.382	N/A	N/A
Metal	0.113	4.624	13.406	37.491	44.366	N/A
Textile	1.639	12.844	52.104	30.993	2.421	N/A
Wood	12.15	0.92	20.688	43.03	11.841	11.376

Table 9: Group's Market Share within Industry. Groups are ordered in increasing mean level of $\hat{\alpha}_{gt}$.

The table 8 shows that each estimated group is generally well populated. So the α_{gt} estimates use ample observations generally across the different groups. Appendix-E has stacked bar plots showing the differences in mean level input choices among the groups. For all sectors, the group's mean level of $\hat{\alpha}_{gt}$ increases with the group's mean level of capital. A costly capital adjustment model can explain this association. To avoid frequently adjusting capital, the firm front-loads its investment needs, and the level of front-loading increases with higher grouped productivity. With more flexible inputs, the firm weighs more on other short term aspects, from the demand side, in its input choices. This aspect explains why the mean level grouped productivity does not have a strict positive relationship with the mean level of intermediate material and labour in the Textile, Wood, and Metal sectors. However, all three inputs hold a fairly positive association with grouped productivity over all sectors.

The table 9 shows that the market shares concentrate within a few groups more than what table 8's population count suggests. The low productivity groups have a disproportionately small market share relative to their firm population. Thus these sectors' aggregate output growths are more sensitive to the productivity changes in the highly productive groups. It may be interesting to match the groups with other observable characteristics to better understand the engine behind the Chilean economic growth in future research.

Output Elasticity Estimates and the Transmission Bias

Sector	$OLS: \hat{\beta}_k$[20]	$OLS: \hat{\beta}_l$	$G=1$:$\hat{\beta}_k$	$G=1$:$\hat{\beta}_l$	$G>1$:$\hat{\beta}_k$	$G>1$:$\hat{\beta}_l$
Food	0.341 (0.016)	0.815 (0.032)	0.33 (0.025)	0.539 (0.045)	0.3 (0.024)	0.517 (0.047)
Metal	0.219 (0.028)	0.917 (0.044)	0.225 (0.041)	0.667 (0.066)	0.151 (0.041)	0.657 (0.068)
Textile	0.233 (0.028)	0.78 (0.044)	0.192 (0.04)	0.72 (0.062)	0.176 (0.042)	0.7 (0.06)
Wood	0.195 (0.036)	0.975 (0.063)	0.156 (0.053)	0.895 (0.095)	0.181 (0.057)	0.829 (0.091)

Table 10: Output Elasticity Estimates - last two columns report the heterogeneous specifications. For the heterogeneous specification: $G = 4$ for Food, $G = 5$ for Metal and Textile, and $G = 6$ for Wood.

Heuristically, the transmission bias is positive for the elasticity estimate of the more flexible input. The firm prefers to adjust for more flexible input when productivity increases. Typically, labour is assumed to be a more flexible input than capital. In line with this theory, the table shows the OLS $\hat{\beta}_l$ is the largest. Then the estimate of β_l further decreases from the single productivity group specification to the heterogeneous productivity group specification.

The heterogeneous groups' elasticity estimates have their return-to-scale hovering between 0.808 and 1.01. They are reasonably close to constant return-to-scale. The reported capital coefficient estimates are statistically significant from zero.[21] These estimates verify the conjecture of grouped fixed effects being more resilient to attenuation bias as compared to fixed effects.[22]

As already mentioned, the information criterion selects the number of groups here. In Appendix E, I plot output elasticity estimates for different G specifications. The output elasticity estimates are quite sensitive over different group specifications. Finding alternative methods to select the number of groups for the production function is an avenue for future research.

Productivity Heterogeneity and Productivity Growth

Here, I assess the difference in productivity measurement from accounting for heterogeneous productivity groups. The first part captures the difference in productivity growth's effect on output. Then the difference in productivity distribution's dispersion is examined.

	$G = 1$	$G > 1$
Food	2.078	2.277
Metal	4.98	5.89
Textile	1.5	1.57
Wood	0.81	0.4

Table 11: Average Output Growth Due to Productivity - Controlling for Inputs Level

Using Olley and Pakes (1996) 's formula, I decompose the annual output growth due to productivity growth after controlling for inputs level. After averaging them over the years, I report them in table 11 for each sector. For the Metal sector, heterogeneous group specification accounts for at least 18 % more growth from productivity - 9.6 % for the Food sector. Interestingly, the Wood sector reports a lower rate under heterogeneity. Appendix E shows that the lowest productivity group experienced a sizeable productivity contraction for some time. Homogeneous specification hides this fact, and it may explain the difference.

[21] The statistical significance is at the 5% level if the bootstrap confidence intervals are valid. The paper has only studied the confidence intervals with 4's simulation.

[22] Lee, Stoyanov, and Zubanov (2019) estimated the output elasticity for the Chilean Textile sector. However, they treated α as a firm fixed effect. Their Textile sector's capital-output elasticity estimate is at a single-digit percentage point and statistically insignificant from zero at the 10% level. The usual suspect is attenuation bias from the capital's measurement error.

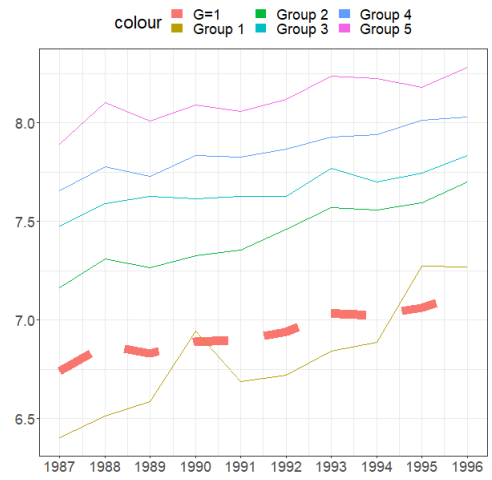


Figure 4: Metal Sector: $\hat{\alpha}_{gt}$'s time-path

Figure 4 shows the homogeneous group specification understates the productivity mean level for most groups in the heterogeneous specification. The graph helps to explain the 18% difference that is documented in table 11. Appendix E has the plots for the other three sectors. Food and Textile sectors also exhibit upward growth trends. The Wood sector's productivity dynamic is more complex and requires more context for interpretation.

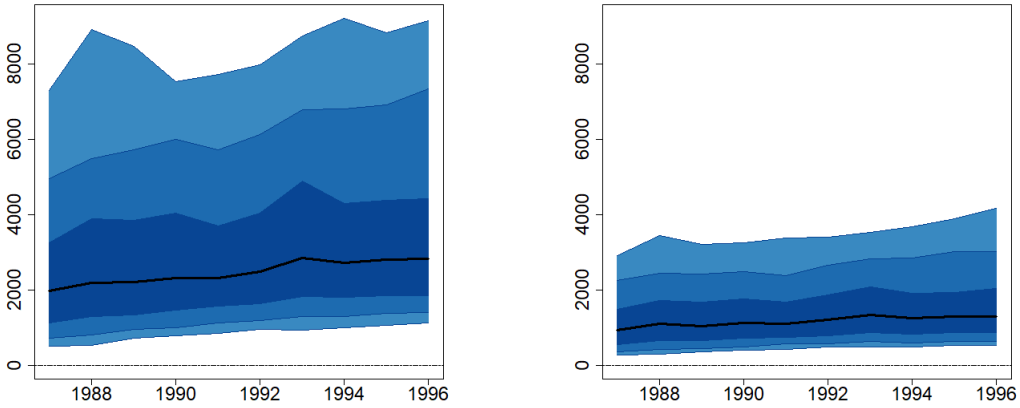


Figure 5: Metal Productivity Fan Charts: 5%,10%,25%,50%,75%,90%,95%. Left: $G = 5$ and Right: $G = 1$.

The graph shows the sectors' weighted[23] productivity distribution from 1987 to 1996. The 75 and 90 percentiles increased twice-fold after accounting for heterogeneous groups. Appendix E presents the fan charts for the other sectors, and the increase in productivity heterogeneity is also noticeable for the Food and Textile sectors.

[23]The weights are the firm's market share in the sector's sample.

6 Conclusion

This paper begins with considering the partially linear semiparametric panel model with additive separable grouped fixed effects and, for the main application, studies its use as the first-step problem in a two-step estimation. The two-step estimation identifies the parameter τ with a general method of moment criterion, conditioning on the partially linear model's parameters.

For the partially linear model, I simultaneously estimate the nonparametric component by the series approach and classify the group memberships by clustering. Subsequently, I propose a consistent estimator of τ based on the sample moments criterion, conditioning on the partially linear model's estimates. For the partially linear model's limit theory, I show the linear coefficient estimator as $\sqrt{NT}$-consistent and asymptotically normal. Furthermore, the grouped fixed effects estimator and nonparametric estimator are uniformly consistent. In the asymptotic limit, the classification achieves the Oracle equivalence, and the linear coefficient estimator is efficient as when the group memberships are known. On estimating the number of groups, I consider the information criterion and show its selection consistency. For the two-step problem, my estimator is consistent.

I use my two-steps problem to extend the proxy variable method, extensively used to estimate the firm's production function. My extension addresses the proxy variable's scalar unobservable problem by introducing firms' productivity as cross-correlated. Now a firm's productivity innovation can have positive spillover effects on other firms. Furthermore, for the intermediate material's structural value-added model, the marginal product of intermediate material can now be time-varying. In Monte Carlo simulation, I find my production function estimator can perform well even under a small T when the groups are well-separated. Furthermore, the information criterion can overwhelmingly select the correct number of groups under a small T.

For the empirical application, I apply my production function estimator on four large Chilean manufacturing sectors from 1987 to 1996. In line with the transmission bias intuition, my estimator downward revises the proxy variable's estimates on the more flexible input's coefficient - the output elasticity for labour. For policy analysis, my analysis shows a significant increase in the productivity distribution's dispersion after introducing heterogeneous productivity groups. Furthermore, productivity also appears more responsible for output growth in the Metal, Food, and Textile sectors.

7 Extension

The paper's main section only considered the time-invariant proxy variable setup. However, it can easily extend to finitely many known structural breaks setup. Formally, there are known breakpoints B_j $(j = 1, ..., J)$ and, between any breakpoints $j-1$ and j, $h_t = h_j$ which satisfies the inversion step. Then the strategy $h_t(k_{it}, l_{it}, \eta_{it}) = \sum_{j \in \{1,...,J+1\}} \{B_{j-1} \leqslant t \leqslant B_j\} h_j(k_{it}, l_{it}, \eta_{it})$, with $B_0 = 1$ and $B_{J+1} = T$ as the convention. In this setup, the partially linear model's modified least squared criterion is

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J+1} \left(y_{it} - x_{it}'\theta - p^K(k_{it}, l_{it}, \upsilon_{it}) \beta^{K,j} - \alpha_{g_i t}^0 \right)^2 \{B_{j-1} \leqslant t \leqslant B_j\}, \tag{35}$$

over $\beta^{K,j}$, α_{gt}, and g_i. The new estimate is

$$\hat{m}_t(k_{it}, l_{it}, \upsilon_{it}) = \sum_{j \in \{1,...,J+1\}} \{B_{j-1} \leqslant t \leqslant B_j\} p^K(k_{it}, l_{it}, \upsilon_{it}) \hat{\beta}^{K,j}. \tag{36}$$

Then in the second step, the moments should be constructed with

$$\eta_t\left(k_{it}, l_{it}, \tau, \{\nu_j\}_{j=1}^{J+1}\right) = \left(\hat{m}_t\left(k_{it}, l_{it}, \upsilon_{it}\right) - f\left(k_{it}, l_{it}; \tau\right) - \nu_j\right)\left\{B_{j-1} \leqslant t \leqslant B_j\right\}. \tag{37}$$

The same set of previous moments should identify τ and ν_j. Even though there are additional ν_j, only one shows up in each moment condition per period.

By adding additional notations, the proofs can be adapted using near-identical main assumptions for all the asymptotic results to hold in this extension. The extension is allowed because the number of non-parametric functions is fixed and not growing with T. Hence, in practice, it is cautious to keep only a few number of breakpoints. Heuristically speaking, over-fitting the number of breakpoints can weaken the data's identification of the group memberships. Then the data may struggle to differentiate the breakpoints' effects separately from the different groups' α_{gt}s.

Stretching the idea further is having the criterion to locate the known J many break points. A naive way to do this is optimizing the modified least-squared criterion also over the B_js. The local optimization algorithm is presented in the next page.

Algorithm 2: Estimating B_j^0 with $\theta^0, \beta^{0,K,j}, g_i^0$, and α_{gt}^0

Initialize $\left\{\hat{g}_{i[0]}\right\}_{i=1}^{N}$ and $\left\{\hat{B}_{j[0]}\right\}_{j=1}^{J}$ subjected to $B_j < B_{j+1}$;

Using $\left\{\hat{g}_{i[0]}\right\}_{i=1}^{N}$ and $\left\{\hat{B}_{j[0]}\right\}_{j=1}^{J}$, estimate $\hat{\theta}_{[0]}, \hat{\beta}_{[0]}^{K,j}$, and $\hat{\alpha}_{gt[0]}$ by minimizing the modified least-squared criterion;

while *convergence is not achieved on the kth iteration* **do**

  By using the kth iteration's $\left\{\hat{B}_{j[k]}\right\}_{j=1}^{J}$, $\hat{\theta}_{[k]}, \hat{\beta}_{[k]}^{K,j}$ and $\hat{\alpha}_{gt[k]}$, update $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^{N}$ to minimize the least squared criterion;

  for i *in 1:J* **do**

    Using $\hat{\theta}_{[k]}, \hat{\beta}_{[k]}^{K,j}$, $\hat{\alpha}_{gt[k]}$, and $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^{N}$, record the modified least squared criterion by moving $\hat{B}_{j[k]}$ up by one period;

    Using $\hat{\theta}_{[k]}, \hat{\beta}_{[k]}^{K,j}$, $\hat{\alpha}_{gt[k]}$, and $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^{N}$, record the modified least squared criterion by moving $\hat{B}_{j[k]}$ down by one period;

  end

  Find the single move leading to the most reduction of the modified least squared criterion within the for-loop;

  Update to $\left\{\hat{B}_{j[k+1]}\right\}_{j=1}^{J}$ by implementing that single move whilst keeping other break points the same;

  Using $\left\{\hat{g}_{i[k+1]}\right\}_{i=1}^{N}$ and $\left\{\hat{B}_{j[k+1]}\right\}_{j=1}^{J}$, estimate $\hat{\theta}_{[k+1]}, \hat{\beta}_{[k+1]}^{K,j}$ and $\hat{\alpha}_{gt[k+1]}$ by minimizing the modified least-squared criterion based;

  Check for convergence of the modified least squared criterion;

end

The asymptotics of this further extension is not covered here.

References

Ackerberg, D., K. Caves, and G. Frazer (2015). "Identification Properties of Recent Production Function Estimators". *Econometrica* 83 (6), pp. 2411–2451.

Ai, C. and X. Chen (2003). "Efficient Estimation of Models with Conditional Moment Restrictions Contains Unknown Functions". *Econometrica* 71, pp. 1795–1843.

Andrews, D. (1983). "First Order Autoregressive Processes and Strong Mixing". *Cowles Foundation Discussion Paper* 664.

Arellano, M. (1987). "Computing Robust Standard Errors for Within-groups Estimators". *Oxford Bulletin of Economics and Statistics* 39, pp. 431–434.

Bai, J. (2009). "Panel Data Models with Interactive Fixed Effects". *Econometrica* 77 (4), pp. 1229–1279.

Bai, J. and T. Ando (2016). "Panel Data Models with Grouped Factor Structure Under Unknown Group Membership". *Journal of Applied Econometrics* 31, pp. 163–191.

Bai, J. and S. Ng (2002). "Determining the Number of Factors in Approximate Factor Models". *Econometrica* 70-1, pp. 191–221.

Banerjee, A. and E. Duflo (2003). "Inequality and Growth: What Can the Data Say?" *Journal of Economic Growth* 8, pp. 267–299.

Blundell, R. and J.L. Powell (Jan. 2000). "Endogeneity in Nonparametric and Semiparametric Regression Models". *Blundell, R. and Powell, J.L. (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont, M. and Hansen, L.P. and Turnovsky, S.J., (eds.) Advances in Economics and Econometrics: Theory and Applications: Eighth World Congress: Volume II. Econometric Society Monographs (36). Cambridge University Press, Cambridge, UK, pp. 312-357. ISBN 9780521818735.*

Bonhomme, S., T. Lamadon, and E. Manresa (2017). "Discretizing Unobserved Heterogeneity". *Working Paper.*

Bonhomme, S. and E. Manresa (2015). "Grouped Patterns of Heterogeneity in Panel Data". *Econometrica* 83 (3), pp. 1147–1184.

Chen, X. (2007). "Large Sample Sieve Estimation of Semi-Nonparametric Models". *Handbook of Econometrics*. Ed. by J.J. Heckman and E.E. Leamer. 1st ed. Vol. 6B. Elsevier. Chap. 76.

Chen, X., O. Linton, and I. Keilegom (2003). "Estimation of Semiparametric Models When the Criterion Function is Not Smooth". *Econometrica* 71, pp. 1591–1608.

Cheng, X., F. Schorfheide, and P. Shao (2019). "Clustering for Multi-dimensional Heterogeneity with Application to Production Function Estimation".

De Loecker, J., J. Eeckhout, and G. Unger (2018). "The Rise of Market Power and the Macroeconomic Implications".

De Loecker, J. and P. Scott (2017). "Estimating market power. Evidence from the US Brewing Industry".

De Loecker, J. and F. Warzynski (2012). "Markups and Firm-level Export Status". *American Economic Review* 102 (6), pp. 2437–2471.

Freyberger, J. (2018). "Non-parametric Panel Data Models with Interactive Fixed Effects". *Review of Economic Studies* 85, pp. 1824–1851.

Gandhi, A., S. Navarro, and D. Rivers (2017a). "On the Identification of Gross Output Production Functions". *University of Western Ontario, Center for Human Capital and Productivity (CHCP) Working Papers* 20181.

— (2017b). "How Heterogeneous is Productivity? A Comparison of Gross Output and Value Added". *University of Western Ontario, Center for Human Capital and Productivity (CHCP) Working Papers* 201727.

Griliches, Z. and J. Hausman (1986). ""Errors in Variables in Panel Data"". *Journal of Econometrics* 31.

Griliches, Z. and J. Mairesse (1998). "Production Functions: The Search for Identification". *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*.

Hansen, C. (2007). "Asymptotic properties of a robust variance matrix estimator for panel data when T is large". *Journal of Econometrics* 141, pp. 597–620.

Huang, X. (2013). "Nonparametric Estimation in Large Panels with Cross-sectional Dependence". *Econometric Reviews* 32 (5-6), pp. 754–777.

Kasahara, H., P. Schrimpf, and M. Suzuki (2017). "Identification and Estimation of Production Function with Unobserved Heterogeneity". *Working Paper*.

Lee, J. and P. Robinson (2016). "Series estimation under cross-sectional dependence". *Journal of Econometrics* 190, pp. 1–17.

Lee, Y., A. Stoyanov, and N. Zubanov (2019). "Olley and Pakes-style Producton Function Estimators with Firm Fixed Effects". *Oxford Bulletin of Economics and Statistics* 81,1, pp. 79–97.

Levinsohn, J. and A. Petrin (2003). "Estimating Production Functions Using Inputs to Control for Unobservables". *Review of Economic Studies* 70 (2), pp. 317–342.

Newey, W. (1997). "Convergence Rates and Asymptotic Normality for Series Estimators". *Journal of Econometrics* 79, pp. 147–168.

Olley, S. and A. Pakes (1995). "A Limit Theorem for a Smooth Class of Semiparametric Estimators". *Journal of Econometrics* 65, pp. 295–332.

— (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry". *Econometrica* 64 (6), pp. 1263–1295.

Qi, L. (2000). "Efficient Estimation of Additive Partially Linear Models". *International Economic Review* 41 (4), pp. 1073–1092.

Robinson, P. (1988). "Root-N-Consistent Semiparametric Regression". *Econometrica* 56, pp. 931–954.

Su, J. and S. Jin (2012). "Sieve Estimation of Panel Data Models with Cross-section dependence". *Journal of Econometrics* 169 (1), pp. 34–47.

Su, L., Z. Shi, and P. Philips (2016). "Identifying Latent Structures in Panel Data". *Econometrica* 6, pp. 2215–2264.