Unobserved Clusters of Time-Varying Heterogeneity in Nonlinear Panel Data Models*

Martin Mugnier†

November 19, 2022

Job Market Paper

Click here to see the latest version.

Abstract

In studies based on longitudinal data, researchers often assume time-invariant unobserved heterogeneity or linear-in-parameters conditional expectations. Violation of these assumptions may lead to poor counterfactuals. I study the identification and estimation of a large class of nonlinear grouped fixed effects (NGFE) models where the relationship between observed covariates and cross-sectional unobserved heterogeneity is left unrestricted but the latter only takes a restricted number of paths over time. I show that the corresponding "clusters" and the nonparametrically specified link function can be point-identified when both dimensions of the panel are large. I propose a semiparametric NGFE estimator and establish its large sample properties in popular binary and count outcome models. Distinctive features of the NGFE estimator are that it is asymptotically normal unbiased at parametric rates, and it allows for the number of periods to grow slowly with the number of cross-sectional units. Monte Carlo simulations suggest good finite sample performance. I apply this new method to revisit the so-called inverted-U relationship between product market competition and innovation. Allowing for clustered patterns of time-varying unobserved heterogeneity leads to a less pronounced inverted-U relationship.

Keywords: nonparametric identification, semiparametric estimation, fixed effects, nonlinear panel data models, clustering, time-varying heterogeneity.

JEL Codes: C14, C23, C25.

*First version: September 30, 2022. I am deeply grateful to my Ph.D. supervisor Xavier D'Haultfœuille for his continuous support and being a source of inspiration for the choice of this topic. Special thanks go to Stéphane Bonhomme, Victor-Emmanuel Brunel, and Ao Wang for invaluable advice and guidance. I am also thankful to Arnak Dalalyan, Laurent Davezies, Jim Heckman, Bo Honoré, Francis Kramarz, Elena Manresa, Isabelle Méjean, Pauline Rossi, Anna Simoni, Anthony Strittmatter, Michael Visser, Andrei Zeleneev, and seminar participants at CREST, the University of Chicago, 2022 EEA-ESEM, 2022 NASMES, and 2021 EWMES conferences for their comments. All remaining errors are my own.

†Ph.D. Candidate in Economics, CREST, ENSAE, Institut Polytechnique de Paris. Email: martin.mugnier@ensae.fr. Website: https://martinmugnier.github.io. This research is supported by several grants of the French National Research Agency (ANR): "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047 and EUR DATA EFM/ANR-18-EURE-0005) and Otelo (ANR-17-CE26-0015-041).

1 Introduction

Unobserved heterogeneity is a prevalent feature of most reduced-form and structural work in economics and other social sciences. Observational outcomes and explanatory variables of interest typically correlate over time with factors unobserved to the researcher. This confounding problem renders identification of key parameters of interest, such as average partial effects, difficult.

By sampling N individuals at T points in time, panel data offer opportunities to account for latent structures embedded in low-dimensional manifolds (see, e.g., Bai, 2009; Wooldridge, 2010; Hsiao, 2015; Moon and Weidner, 2019; Bonhomme, Lamadon, and Manresa, 2022).[1] While random effects approaches specify the conditional distribution of the unobserved heterogeneity given covariates (up to a few parameters), fixed effects approaches leave this distribution unrestricted and introduce instead many additional parameters. In particular, pooled linear regression with additively separable individual and time-specific effects has been widely used to control for unobserved permanent heterogeneity and "common trends". de Chaisemartin and D'Haultfœuille (2020) find that 20% of applied papers published in the *AER* between 2010-12 have estimated such a regression.

The underlying two-way fixed effects model, however, is restrictive in at least two important ways. First, it cannot accomodate nonlinearity and nonseparability in parameters that frequently arise from economic theory and de facto imply heterogeneous partial effects (e.g., discrete choice, point mass in outcome). Second, common trend assumptions may fail (see, e.g., Roth and Rambachan, 2022) and the model does not capture more complicated patterns of time-varying unobserved heterogeneity.

Jointly adressing these concerns is difficult. Standard differencing techniques or sufficient statistics for the unobserved effects are generally lacking in nonseparable models. Allowing for unobserved diverging trends creates a dimensionality challenge in identification and estimation, which reflects Neyman and Scott (1948)'s well-known incidental parameters problem (even with large T).

Among existing approaches, restricting the support of unobserved heterogeneity has recently gained increasing attention as an interpretable, flexible, and economically meaningful dimension-reduction device.[2] Specifically, it often is plausible that individuals partition into a moderate number of clusters such that all cluster members share the same path of unobserved heterogeneity over time but the partition is unknown to the researcher. The problem becomes that of classifying a large number of individuals into clusters and estimating a large number of nonseparable cluster-specific time effects in "large-N, T" nonlinear panel models, where N and T jointly diverge to infinity.[3]

To the best of my knowledge, no result is known concerning the nonparametric identification of many nonlinear models widely used in empirical research (e.g., random utility binary/ordered choice

[1]This echoes Occam's razor principle and the "manifold hypothesis" (Goodfellow, Bengio, and Courville, 2016).

[2]Pioneering work includes Heckman and Singer (1984); Hahn and Moon (2010); Bonhomme and Manresa (2015).

[3]Such asymptotics have become increasingly popular in the last decades, given the growing availability of high-frequency data (e.g., scanner, financial data). See, among others, Hahn and Newey (2004); Arellano and Hahn (2007); Dhaene and Jochmans (2015); Fernández-Val and Weidner (2016); Chen, Fernández-Val, and Weidner (2021); Chen, Rysman, Wang, and Wozniak (2022).

models) in this setting.[4] Furthermore, estimation and inference using recently proposed semiparametric estimators (e.g., interactive fixed effects) is quite challenging. Asymptotic distributions are rarely available or must be bias-corrected using analytical or jackknife methods justified by asymptotic frameworks where N and T grow at the same rate (see Zeleneev, 2020; Chen, Fernández-Val, and Weidner, 2021; Bonhomme, Lamadon, and Manresa, 2022). This gap in identification and these limitations in semiparametric estimation are important. The distribution of idiosyncratic error terms (e.g., random shocks in taste), together with common and fixed effects parameters, is a building block for estimating counterfactual events and policy-relevant parameters such as average causal effects. Unjustified parametric assumptions can be expected to deliver poor counterfactuals in large panels. Also, T is often much smaller than N in practice.

In this paper, I address both of these concerns for a large class of nonseparable nonlinear grouped fixed effects (NGFE hereafter) single-index static models for discrete outcomes. In the most simple version, individual $i \in \{1, ..., N\}$'s outcome $Y_{it} \in \mathcal{Y}$ at time period $t \in \{1, ..., T\}$ given i's covariates history $X_{i1}, ..., X_{it}$, cluster membership $g_i \in \{1, ..., G\}$, and cluster-specific effect $\alpha_{g_i t}$ is such that

$$\Pr\left(Y_{it} = y \mid X_{i1}, ..., X_{it}, g_i, \alpha_{g_i t}\right) = h\left(y, X_{it}'\beta + \alpha_{g_i t}\right), \tag{1}$$

where the common parameter β, the link function $h(\cdot, \cdot)$, the number of clusters $G \ll N$, the cluster memberships g_i, and cluster-specific effects $(\alpha_{gt})_{g,t}$ are unobserved to the econometrician and treated as parameters to estimate. This class covers many important models of empirical interest such as binary choice, ordered choice, and count data (see Section 2). Extensions to multinomial choice or fully nonparametric models are discussed in the Appendix.

In this context, I make two contributions. My first contribution is to provide primitive conditions under which all parameters of model (1) are point identified as N and T grow large. The proof is constructive and relies on two steps. In a first step, I draw on an injectivity condition à la Bonhomme, Lamadon, and Manresa (2022) (see their Assumption 2) to build test functions which identify the sequence of latent clusterings $\{g_1, ..., g_N\}_{N \geq 1}$ and number of clusters G from pairwise comparisons of conditional probability functions identified by time variations in the data at the individual level. The key idea is to circumvent the difficult (nonlinear and NP-hard) k-means clustering problem by considering instead $N(N-1)/2$ individuals-pairing testing problems.[5] I show that the injectivity condition holds if, for instance, clusters are "well separated", there is continuous local variation in a "special" regressor (not necessarily with large support), and the link function is real-analytic (see, e.g., Krantz and Parks, 2002).[6] In a second step, I resort to within-cluster variation and apply a well-known result by Ichimura (1993) to identify the common slope

[4] While Fernández-Val and Weidner (2018) argue "*most models are point identified with large T*", this paper gives sufficient conditions for a large class of models.

[5] This idea is at the core of many "hierarchical" or "agglomerative" approaches proposed in the unsupervised learning literature (e.g., DBSCAN clustering algorithm).

[6] Special regressors are widely used in econometrics (for discussion and examples see, e.g., Lewbel, 2014). There is a trade-off between imposing (i) analyticity of the link function which allows to interpolate from bounded variation in the regressors at the cost of a strong functional form assumption and (ii) the existence of a special regressor with unbounded support.

parameter β up to scale. Identification of cluster-specific time-varying effects and the unknown link function then follows from leveraging compensating variations within and between clusters and a monotonicity property.

My second contribution is to develop simple NGFE semiparametric estimators and establish their large sample properties.[7] I introduce a general M-estimation framework to estimating nonlinear models with clusters of time-varying unobserved heterogeneity. Semiparametric NGFE estimators are obtained by specializing the framework to models with a known link function and a known number of clusters. These estimators maximize the likelihood of the data conditional on the latent clustering and time-effects. Importantly, no tuning parameter is required. However, computation can be cumbersome in large samples. In a companion paper (Mugnier, 2022), I show how combining nuclear norm regularization with the pairwise differencing argument that serves as a foundation for the present nonparametric identification analysis delivers a computationally trivial estimator for a linear version of model (1), which enjoys more powerful statistical guarantees than Bonhomme and Manresa (2015)'s grouped fixed effect estimator. In particular, the unknown number of clusters G can be consistently estimated under a restricted eigenvalue condition and without prior knowledge of an upper bound $G_{\max} \geq G$ (see Proposition 3.1 in Mugnier, 2022). Here, I instead propose a simple heuristic, namely Lloyd (1982)'s algorithm described in Section 4.3, and show that it performs well in various Monte Carlo experiments with moderate sample sizes and number of clusters (see Section 6). From a theoretical viewpoint, and in contradistinction with popular fixed effects estimators such as Chamberlain (1980); Rasch (1960) or Charbonneau (2017)'s conditional logit, NGFE estimators can accomodate time-invariant regressors and do not drop individuals without any variation in outcomes, thus exploiting the full sample variation. On the other hand, contrary to Bonhomme, Lamadon, and Manresa (2022), I maintain the assumption that unobserved heterogeneity is discrete. This assumption is key for the NGFE estimator to have only one optimization step and enjoy a "perfect recovery" property: provided T grows at least as some power of N, the misclassification probability tends to zero uniformly across individuals.[8] As in the linear case (see Bonhomme and Manresa, 2015), this result implies that, under additional regularity conditions, NGFE estimators of the slope and cluster-specific effects are asymptotically equivalent to the infeasible oracle maximum likelihood estimator (MLE) based on knowledge of the clustering. Remarkably, when $T = o(N)$, this oracle is asymptotically unbiased so that standard MLE inference yields tests and confidence intervals with correct asymptotic level. When $N/T \to \kappa \in (0, +\infty)$, existing results can be applied to the oracle to derive analytical or jackknife bias correction methods for the slope and average marginal effects estimates.[9]

I investigate the finite sample performance of NGFE estimators, as well as large-N, T estimators of their variance, by means of Monte Carlo simulations. I compare the results with state-of-the-art competing methods. I find that NGFE estimators perform quite well in settings they are

[7] Fully nonparametric estimation could follow the constructive identification argument. I do not pursue this avenue here because it would require a lot of tuning parameters.

[8] A concentration inequality for martingale differences due to Lesigne and Volný (2001) is used to show this result.

[9] See, e.g., Hahn and Newey (2004), Arellano and Hahn (2007), and Chen, Fernández-Val, and Weidner (2021).

meant for. In particular, in a static logit model with clustered time-varying correlated unobserved heterogeneity, $N = 90$, $T = 7$, the NGFE estimator has the smallest bias and Root Mean Square Error (RMSE) compared to both linear and nonlinear methods such as linear two-way fixed effects (TWFE), Bonhomme and Manresa (2015)'s grouped fixed effects (GFE) Bonhomme, Lamadon, and Manresa (2022)'s 2-step GFE, Fernández-Val and Weidner (2016)'s nonlinear TWFE, or Rasch (1960); Chamberlain (1980)'s conditional MLE. It also has the best finite-sample 95% CI's coverage (84 to 86%) compared to the CMLE (less than 50%).[10] It takes 10 seconds to compute on a professional laptop, which is similar to that of competing clustering methods such as 2-step GFE.

Finally, I illustrate the practical usefulness of NGFE estimators by revisiting an influential paper by Aghion, Bloom, Blundell, Griffith, and Howitt (2005). The authors investigate the relationship between product market competition and innovation using a panel of seventeen UK industries (i) that spans the last part of the twentieth century ($t = 1973, ..., 1994$). Their preferred specification is a nonlinear Poisson regression model of the number of citation-weighted patents on "one minus the Lerner index" that controls for multiplicatively separable industry and time effects. Their results suggest a strong inverted-U relationship. Yet, there is no reason a priori to assume that dynamic shocks driving both the production of patents and the market structure of industries are common to all industries. When I estimate a NGFE model, I find a much flatter inverted-U curve. This is due to the presence of clustered patterns of time-varying unobserved heterogeneity. The data-driven clustering procedure reveals a permanently "high (resp. low)-innovation" cluster of industries gathering "heavy (resp. light) sectors" such as automobile production, chemical products (resp. manufacture of paper/paper products, textile industry), as well as transitioning "caching-up" clusters of industries, including data and tech related sectors such as electrical and electronic engineering or data processing equipment. These new results shed light on unobserved diverging mechanisms that drive both the market structure and technological change across time. Cluster memberships and clusters effects can be further used as dependent variables to guide the search of key time-varying omitted variables determining both technological change and market structure.

Economics provides many other possible applications of NGFE models. Janys and Siflinger (2021) find that young women engage into systematically divergent unobserved risky behaviors over time that simultaneously affect the chance to have an abortion and that to develop mental health disorders (a binary dependent outcome in the study). Deb and Trivedi (1997) control for unobserved time-invariant discrete types of health risk. More generally, any limited dependent variable model (e.g., ordered, multinomial logit) in which it is expected that the baseline level of cross-sectional unobserved heterogeneity is not subject to the same trend across individuals (e.g., human capital accumulation, change in taste for different products in the long run) is a candidate. The approach could also be applied to network data with clustered patterns of heterogeneity (e.g., gravity equations in trade), which I leave for further research (see Section B.3).

Overall, the theoretical results broaden the scope of application of GFE estimators and clustering

[10]Note that only Bonhomme, Lamadon, and Manresa (2022)'s estimator assumes a correctly specified model. This paper does not provide inference tools. Comparison with Chen, Fernández-Val, and Weidner (2021)'s estimator is left for further research.

techniques in econometrics, complementing the available toolbox for applied economists interested in assessing the robustness of their results to specification choices. Results from the empirical applications confirm the usefulness of considering flexible specifications such as NGFE for modeling unobserved heterogeneity.

This paper contributes to the large literature on nonseparable panel data models. Most previous papers from this literature obtain (partial) identification results under fixed-T asymptotics. Point-identification results with fixed T are scarce, even in a static simple binary choice model with unit-specific unobserved effect (see, e.g., Chamberlain, 2010; Davezies, D'Haultfoeuille, and Mugnier, 2020). Some papers have leveraged the large-T dimension but otherwise rely on (additive) separability of the individual/time unobserved heterogeneity or parametrically specify the link function.[11] In contrast, I alleviate the large-T dimension, cluster separation, and the single-index clustered structure to show that all parameters of NGFE models can be (nonparametrically) point-identified even with clustered time patterns of unobserved heterogeneity. I use the technique of *compensating variations* like D'Haultfoeuille, Hoderlein, and Sasaki (2021) and Mugnier and Wang (2022), which does not necessarily require large support (see also Vytlacil and Yildiz, 2007). This paper also contributes to the literature estimating semiparametric nonlinear large-N, large-T panel data models with multiple fixed effects. Much previous work in the panel data literature has focused on estimation of semiparametric factor-analytic type linear models while nonlinear models with interactive fixed-effects have only recently drawn considerable attention.[12] Fernández-Val and Weidner (2016), Graham (2017), and Charbonneau (2017) provide consistent and asymptotically normal semiparametric estimators of the homogeneous slope coefficient (as well as average partial effects in Fernández-Val and Weidner, 2016) in nonlinear TWFE models. In contrast to NGFE estimators, Graham (2017) and Charbonneau (2017)'s conditioning estimators, by partialling out unobserved effects, do not provide consistent estimates for them, and Fernández-Val and Weidner (2016) require $N/T \to \kappa \in (0, +\infty)$ to obtain statistical guarantees. Neither TWFE nor NGFE models are nested one into another and the two approaches should therefore be seen as complementary. On the other hand, some papers assume that clusters are known to the econometrician (see, e.g., Bester and Hansen, 2016; Arkhangelsky and Imbens, 2018). Many papers allow for a latent clustered structure but otherwise impose time-invariant or additively separable unobserved heterogeneity.[13] Differently from us, a line of research put the grouping assumption on the unknown slope coefficient (heterogeneous models), letting again the unobserved heterogeneity individual-specific and time-constant.[14] Allowing for clustered patterns of time-varying unobserved heterogeneity in nonlinear models seems

[11] See, e.g., Vogt and Linton (2017); Zeleneev (2020); Mugnier and Wang (2022).

[12] For linear factor-type models, see, among many others, Bai (2003); Pesaran (2006); Bai (2009); Bonhomme and Manresa (2015); Moon and Weidner (2015); Ke, Li, and Zhang (2016); Moon and Weidner (2017); Ando and Bai (2017). For nonlinear ones, see, e.g., Chen, Fernández-Val, and Weidner (2021); Bonhomme, Lamadon, and Manresa (2022); Ando and Bai (2022).

[13] See, e.g., Bryant and Williamson (1978); Hahn and Moon (2010); Saggio (2012); Bonhomme and Manresa (2015); Su, Shi, and Phillips (2016); Vogt and Linton (2017); Gu and Volgushev (2019); Cheng, Schorfheide, and Shao (2021); Yu, Gu, and Volgushev (2022).

[14] See, Boneva, Linton, and Vogt (2015); Su, Shi, and Phillips (2016); Su, Wang, and Jin (2019); Zhang, Wang, and Zhu (2019); Gao, Xia, and Zhu (2020); Liu, Shang, Zhang, and Zhou (2020); Wang and Su (2021).

to be a difficult and much less investigated problem that I address in this paper. The closest papers to ours are Chen, Fernández-Val, and Weidner (2021), Bonhomme, Lamadon, and Manresa (2022), and Ando and Bai (2022). Chen, Fernández-Val, and Weidner (2021) extend Fernández-Val and Weidner (2016)'s results to semiparametric nonlinear factor-analytic models under concavity conditions. When the link function is parametrically specified, NGFE models are special cases of their framework. In contrast, I consider an unknown link function, derive more primitive conditions for point identification (e.g., monotonicity in place of log-concavity of the MLE), and allow T to grow slowly with N in estimation. The two-step discretization approach developed in Bonhomme, Lamadon, and Manresa (2022), albeit its remarkable generality, comes at a similar price. When heterogeneity is discrete, it ressembles a Lloyd's algorithm where the first clustering step would not take advantage of improvement on the other parameters. Yet, in contrast to the NGFE approach and to the best of my knowledge, no inference result is known for this method. Independently from this paper, Ando and Bai (2022) generalize Bonhomme and Manresa (2015)'s semiparametric GFE estimator to an exponential family of nonlinear grouped factor models with heterogeneous coefficients (including Probit, Logit, Poisson). As in this paper, they consider the MLE and their results allow for heterogeneous coefficients. But their general framework imposes stronger restrictions (requires larger T in the asymptotics), delivers $\sqrt{T}$-rate for the slope coefficient estimates (v.s. $\sqrt{NT}$ for the NGFE estimate of the common slope), and they do not provide nonparametric identification results. A third strand of literature this paper contributes to is that of dimension reduction methods applied to nonlinear panel data models. A surge of papers have leveraged state-of-the-art statistical learning tools such as matrix completion devices and extensions of Tibshirani (1996)'s Least Absolute Shrinkage Estimator (LASSO) estimator to tackle the problem of estimating a large number of unobserved effects in parsimonious panel data models.[15] A common unifying idea is to exploit restrictions on the support of the unobserved heterogeneity, which echoes the concept of sparsity in high-dimensional statistics,[16] or (nonparametric) finite mixtures models and clustering (see, e.g., Forgy, 1965; MacQueen, 1967; Lloyd, 1982; McLachlan and Peel, 2000).

In Section 2, I introduce the class of NGFE models. The main identification result is presented in Section 3. In Section 4, I propose a general M-estimation framework, develop semiparametric NGFE estimators, and discuss computational aspects. Section 5 provides large sample properties in semiparametric binary choice models. Section 6 presents Monte Carlo results. Section 7 contains the empirical application. Section 8 concludes. All proofs are collected in the appendix. For any set A, I let $A^* := A \setminus \{0\}$ and $|A|$ denote the cardinal of A. Henceforth, I denote by $\mathrm{Supp}(U)$ the support of any random variable U.

[15]See, among others, Kock (2016); Kock and Tang (2019); Moon and Weidner (2019); Zeleneev (2020); Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021).

[16]See, e.g., the monograph by Giraud (2014) for a thorough introduction to the topic.

2 Nonlinear Discrete Outcome Models With Unobserved Clusters of Time-Varying Heterogeneity

Suppose to observe a random sample of balanced panel data $\{(Y_{it}, X_{it}')' : (i,t) \in \mathcal{N} \times \mathcal{T}\}$ with dimensions $N := |\mathcal{N}|$ and $T := |\mathcal{T}|$.[17] In many applications, $\mathcal{N}$ is an index for individuals or "units", and $\mathcal{T}$ indexes time periods or "unit members". I consider the problem of modeling, for individual $i \in \mathcal{N}$, the T-vector of discrete outcomes $Y_i = (Y_{it})'_{t\in\mathcal{T}}$ in relation with its $T \times p$ matrix of weakly exogeneous covariates $X_i = (X_{it}')'_{t\in\mathcal{T}}$. The dependent variable Y_{it} represent agents' (choice) decisions and X_{it} represent agents' attributes over time and it is often plausible that time-varying unobservables (to the econometrician) confound the "effect" of X_{it} on Y_{it}.[18]. For instance, in the empirical application, $Y_{it} \in \mathbb{N}$ denotes the number of patents produced by industry i at time t and X_{it} collects industry i's characteristics at time t such as the level of product market competition.

With this purpose, I introduce below a class of nonlinear clustered or "grouped" fixed effects (NGFE) models to flexibly incorporate time-varying patterns of unobserved heterogeneity. I let $\text{Supp}(Y_{it}, X_{it}) = \mathcal{Y} \times \mathcal{X}_i$ and assume that $\mathcal{Y} \subset \mathbb{R}$ is at most countable and $\mathcal{X}_i \subset \mathbb{R}^p$ for some fixed $p \in \mathbb{N}^*$. I assume that individual $i \in \mathcal{N} := \{1, ..., N\}$ at time $t \in \mathcal{T} := \{1, ..., T\}$ chooses $Y_{it} \in \mathcal{Y}$ given her weakly exogeneous covariates $X_i^t := (X_{i1}', ...X_{it}')'$, her unobserved cluster membership variable $g_i^0 \in \mathcal{G}^0 := \{1, ..., G^0\}$, and unobserved time-varying cluster-specific effect $\alpha_{gt}^0 \in \mathcal{A} \subset \mathbb{R}$ such that, for all $y \in \mathcal{Y}$,

$$\Pr\left(Y_{it} = y \mid X_{i1}, ..., X_{it}, g_i^0, \alpha^0_{g_i^0 t}\right) = h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right), \tag{2}$$

where $\beta^0 \in \mathcal{B} \subset \mathbb{R}^p$ in an unknown fixed parameter of interest, $G^0 \in \mathbb{N}^*$ is unknown but "small" relative to N, and $h^0 \in \mathcal{H}$ is an unknown link function from the set

$$\mathcal{H} \subset \left\{ h : \mathcal{Y} \times \mathbb{R} \to (0,1) \text{ measurable}, \sum_{y\in\mathcal{Y}} h(y,\cdot) = 1, \text{ and } \sum_{y\in\mathcal{Y}} |y|\, h(y,\cdot) < \infty \text{ a.e.} \right\}.$$

The common parameter β^0 is often of key interest in applications (e.g., ratios of marginal utilities). For $g \in \mathcal{G}^0$, unobserved cluster-specific time effects $(\alpha^0_{g^0 t})_{t\geq 1}$ account for time-varying unobserved heterogeneity shared by all members of cluster g, i.e., all individuals from the set $\left\{j \in \mathcal{N} : g_j^0 = g\right\}$. These effects are treated as fixed in the analysis but might be arbitrarily correlated with X_{it} and confound β^0. The contemporaneous covariates X_{it} and the unobserved effect $\alpha^0_{g_i^0 t}$ enter the response function as the combination of a linear single-index $X_{it}'\beta^0 + \alpha^0_{g_i^0 t}$ and an unknown

[17] Unbalanced panels can be accomodated easily under exogeneous attrition (i.e., missing-at-random). Endogeneous attrition is beyond the scope of this paper. Throughout the main text, I rule out undirected graph (or network or "pseudo-panel") data for which there is no proper $\mathcal{T}$ and observations are indexed by pairs of indices $(i,t) \in \mathcal{N}^2$ such that $(Y_{it}, X_{it}')' = (Y_{ti}, X_{ti}')'$ for all $(i,t) \in \mathcal{N}^2$. There is a vast literature on models of link formations and networks (see, e.g., de Paula, 2020, for a recent review). I discuss a particular case in Appendix B.3.

[18] E.g., agents choose X_{it} depending on time-varying unobservables that also affect Y_{it} before idiosyncratic shocks are realized. One might also want to distinguish between state dependence and unobserved (time-varying) heterogeneity (see, e.g. Heckman, 1981).

link function h^0.[19] Single index assumptions are common in the nonseparable panel data models literature and serve mainly computational and interpretation purposes (relying on another smooth index would not significantly change our subsequent results, but likely some identification assumptions). The link function h^0 may encapsulate the conditional distribution of random idiosyncratic shocks in latent variable utility choice models with exogeneous covariates. Note that (i) neither the clustering nor the number of clusters is observed by the econometrician and (ii) the number of possible assignments of N individuals into G^0 clusters grows exponentially fast with N.

Model (2), although static (h is not indexed by time), complements models with additively separable (and time-invariant) fixed effects that have been routinely employed in the empirical microeconometric, industrial organisation, macroeconometric, innovation, and international trade literature. I provide below some leading examples.

Example 1 (Binary outcome)

$$Y_{it} = \mathbf{1}\left\{X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} \geq 0\right\},$$

where ε_{it} *is independent from* $\left(X_{i1}', ..., X_{it}', g_i^0, \alpha^0_{g_i^0 t}\right)'$ *and distributed with (unknown) cumulative distribution function (cdf)* Ψ^0. *Then,*

$$h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) = \mathbf{1}\{y=1\} \times \Psi^0\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) + \mathbf{1}\{y=0\} \times \left[1 - \Psi^0\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right)\right].$$

Example 2 (Ordered outcome)

$$Y_{it} = \begin{cases} 0 & \text{if } X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} < d_1^0. \\ 1 & \text{if } d_1^0 \leq X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} < d_2^0. \\ 2 & \text{if } X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} \geq d_2^0, \end{cases} \tag{3}$$

where $d_2^0 > d_1^0$, *and* ε_{it} *is independent from* $\left(X_{i1}', ..., X_{it}', g_i^0, \alpha^0_{g_i^0 t}\right)'$ *and distributed with (unknown) cdf* Ψ^0. *Then,*

$$h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) = \begin{cases} 1 - \Psi^0\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - d_1^0\right) & \text{if } y = 0. \\ \Psi^0\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - d_1^0\right) - \Psi^0\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - d_2^0\right) & \text{if } y = 1. \\ \Psi^0\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - d_2^0\right) & \text{if } y = 2. \end{cases}$$

Example 3 (Count outcome) $\mathcal{Y} = \{0, 1, 2, ...\}$. *A Poisson parametrization specifies*

$$h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) = \frac{\left(\lambda_{it}^0\right)^y \exp\left(-\lambda_{it}^0\right)}{y!}, \tag{4}$$

[19]If h^0 were known to the econometrician, model (2) would become a special case of the semiparametric nonlinear factor models considered in Chen, Fernández-Val, and Weidner (2021).

where $\lambda_{it}^0 = \exp\left(X_{it}'\beta^0 + \alpha_{g_i^0 t}^0\right)$. *Alternatively,* h^0 *could encapsulate, e.g., the negative binomial distribution.*

I adopt the so-called "fixed effects" approach, treating realizations of the unobserved time effects and group membership variables as unrestricted parameters to be estimated. I assume that G^0 is fixed and exogeneous. Policy parameters of interest such as average marginal effects often write as functionals of β^0, h^0, $\alpha^0 := (\alpha_{11}^0, ..., \alpha_{1T}^0, ..., \alpha_{G^01}^0, ..., \alpha_{G^0T}^0)' \in \mathcal{A}^{G^0T}$, and latent clustering structure $\gamma^0 := (g_1^0, ..., g_N^0)' \in \mathcal{G}^{0N}$. Hereafter, I focus on identification and estimation of the sequence of parameters $\theta_{NT}^0 := \left(G^0, h^0, \beta^{0\prime}, \gamma^{0\prime}, \alpha^{0\prime}\right)' \in \Theta_{NT}$, where I let

$$\Theta_{NT} = \bigcup_{G=1}^{+\infty} \{G\} \times \mathcal{H} \times \mathcal{B} \times \{1, ..., G\}^N \times \mathcal{A}^{GT}.$$

While $\mathcal{B}$ is a finite-dimensional space, $\mathcal{H}$ is clearly not and the dimensions of both the discrete set $\{1, ..., G\}^N$ and $\mathcal{A}^{GT}$ grow with the sample size. This makes model (2) a high-dimensional combinatorial semi-parametric nonseparable model.

Remark 1 *It is straightforward to adapt the analysis to allow for cluster-specific slope coefficient* $\boldsymbol{\beta}^0 := \left(\beta_1^{0\prime}, ..., \beta_{G^0}^{0\prime}\right)'$ *such that*

$$\Pr\left(Y_{it} = y \mid X_{i1}, ..., X_{it}, g_i^0, \alpha_{g_i^0 t}^0, \beta_{g_i^0}^0\right) = h^0\left(y, X_{it}'\beta_{g_i^0}^0 + \alpha_{g_i^0 t}^0\right), \quad \forall y \in \mathcal{Y}. \tag{5}$$

I discuss this extension, as well as heterogeneous link functions, additional individual- and time-specific effects, and grouped time-periods in Appendices B.1-B.3. Model (2) can also be extended to allow for multimodal outcomes. The notation are more lengthy and would essentially follow the same lines as in Mugnier and Wang (2022).

Remark 2 *Model (2) extends Bonhomme and Manresa (2015) to nonparametric discrete choice modeling. In contrast to Bonhomme, Lamadon, and Manresa (2022), the link function* h^0 *is unknown, the true underlying unobserved heterogeneity is discrete, and all parameters of the models are considered as target parameters.*

3 Large-N, Large-T Nonparametric Identification

In this section, I prove the nonparametric identification of θ_{NT}^0 in model (2) as N and T diverge jointly to infinity. More specifically, I show that all parameters can be written as known functions of quantities that are point identified from either or both the cross-sectional and longitudinal variation in the data. Note that model (2) is related to nonseparable panel data models with latent factors as it implies the following semiparametric regression equations:

$$\mathbf{1}\{Y_{it} = y\} = h^0\left(y, X_{it}'\beta^0 + \alpha_{g_i^0 t}^0\right) + \varepsilon_{it}(y), \quad \forall (i, t, y) \in \times \mathcal{N} \times \mathcal{T} \times \mathcal{Y}, \tag{6}$$

where $\mathbb{E}\left[\varepsilon_{it}(y) \mid X_i, g_i^0, \alpha^0_{g_i^0 t}\right] = 0$, and

$$Y_{it} = \sum_{y \in \mathcal{Y}} y h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) + v_{it}, \quad \forall (i,t) \in \times \mathcal{N} \times \mathcal{T}, \tag{7}$$

where $v_{it} = \sum_{y \in \mathcal{Y}} y \varepsilon_{it}(y)$ and, by linearity, $\mathbb{E}\left[v_{it} \mid X_i, g_i^0, \alpha^0_{g_i^0 t}\right] = 0$. The representation given by (6) is useful to identify the clustering structure, while the representation given by (7) allows to apply results in Ichimura (1993) under appropriate dependence conditions that I now introduce.

Since both g_i^0 and $\alpha^0_{g_i^0 t}$ are unobserved, identification holds up to clusters relabeling only.[20] It is also necessary to impose location and scale normalizations, which I specify as $\left\|\beta^0\right\| = 1$ and $\alpha^0_{11} = 0$, where $\left\| \cdot \right\|$ denotes the Euclidean norm.[21] Identification is based on Assumptions 1-5 below.

Assumption 1 (Random sampling) *There exist random vectors of fixed dimensions λ^0_{gt}, μ^0_g, ξ^0_i such that, letting $\lambda^0 := \left\{\lambda^0_{gt} : (g,t)\right\}$, $\mu^0 := \left\{\mu^0_g : g\right\}$, $\xi^0 := \{\xi^0_i : i\}$:*

(a) $(Y_i', X_i', g_i^0)'$ is i.i.d. across $i \in \mathcal{N}$ conditional on $\{\alpha^0, \lambda^0, \mu^0\}$.

(b) For all $i \in \mathcal{N}$: $\left(Y_{it}, X_{it}', \alpha^0_{g_i^0 t}\right)'_{t \geq 2}$ is a strictly stationary strong mixing process with mixing coefficients $\tau_i(\cdot)$ conditional on $g_i^0, \mu^0_{g_i^0}, \xi^0_i$. Let $\tau(\cdot) = \sup_i \tau_i(\cdot)$ satisfy $\tau(l) \leq Cm^l$ with $C > 0$ and $m \in (0,1)$.

(c) For all $t \in \mathcal{T}$: $Y_{1t} \mid X_{1t}, g_1^0, \alpha^0, \lambda^0, \mu^0, \xi^0 \overset{d}{=} Y_{1t} \mid X_{1t}, g_1^0, \alpha^0_{g_1^0 t}$.

Assumptions 1(a)-1(b) restrict cross-sectional and time dependence in the data. They allow for flexible patterns of unconditional spatial and time-series correlations as captured by the clustering latent structure $\alpha^0, \lambda^0, \mu^0$ and individual-specific effects ξ^0. Assumption 1(c) requires that λ^0, μ^0, ξ^0 have no effect on the outcome after conditioning for the covariates, cluster membership and the cluster-specific effects α^0. In Appendix B.1, I discuss several extensions such as cluster-specific slopes, individual-fixed and time-fixed effects which possibly affect all observed variables.[22]

Assumption 2 (Latent clustering) *$\mathcal{X} := \bigcap_{i=1}^{\infty} \mathcal{X}_i$ is not empty and:*

(a) There exist known $\mathcal{X}^0 \subset \mathcal{X}$, $y \in \mathcal{Y}$, and functional ϕ such that, for all fixed $(i,j) \in \mathcal{N}^2$, letting $\rho_i(x) : \mathcal{X}^0 \ni x \mapsto \Pr\left(Y_{i2} = y \mid X_{i2} = x, g_i^0, \mu^0_{g_i^0}, \xi^0_i\right)$, $\phi(\rho_i, \rho_j) = \mathbf{1}\left\{g_i^0 = g_j^0\right\}$.

(b) For all $g \in \mathcal{G}^0$, almost surely $\Pr\left(g_1^0 = g \mid \alpha^0, \lambda^0, \mu^0, \xi^0\right) > 0$.

Assumption 2(a) requires clusters to be sufficiently well-separated in terms of individual-level conditional probability functions. It is a low-level injectivity or "completeness"-type assumption à la Bonhomme, Lamadon, and Manresa (2022) which ensures that latent variables are recoverable from

[20] This mirrors rotational invariance normalizations in interactive fixed effects models (see, e.g., Bai, 2009).

[21] These choices are, of course, arbitrary but normalizing $\left\|\beta^0\right\| = 1$ is standard in nonparametric single-index models (see, e.g. Ichimura, 1993; Botosaru and Muris, 2017).

[22] In some application, it could be useful to allow for a non-scalar α^0_{gt}. Estimation in semiparametric nonlinear grouped factor models with many factors has recently been considered in Ando and Bai (2022).

observed moments and leaves flexibility to the researcher for defining clusters of unobserved heterogeneity. In Appendix A.2, I provide sufficient conditions for Assumption 2(a) to hold, including smoothness and the existence of a special regressor à la Honoré and Lewbel (2002) but (possibly) without large support. For such a mapping to exist, the intuition is that whenever $g_i^0 \neq g_j^0$, the conditional distributions $\alpha_{g2}^0 \mid X_{i2} = x, g_i^0, \mu_i^0, \xi_i^0$ and $\alpha_{\tilde{g}2}^0 \mid X_{j2} = x, g_j^0, \mu_j^0, \xi_j^0$ across $x \in \mathcal{X}^0$ should differ sufficiently (and the link function h^0 should be sufficiently smooth to convey such a difference) so as to trigger a difference in the integrated-out conditional outcome probabilities captured by ϕ. In many application, $\phi(f,g) = \mathbf{1}\{f = g\}$ makes sense (see, e.g., Vogt and Linton, 2017). Yet, the setting is kept slightly more general as other clustering structures might be plausible. Assumption 2(b) rules out asymptotically negligible clusters. Notice that allowing for an increasing number of clusters or negligible clusters would need substantial changes to Assumption 1 (e.g., as the cross-sectional identical distribution would not hold anymore). Note also that Assumption 2(a) could be generalized to be based instead on the (possibly infinite dimensional) full conditional distribution of the outcome.

Assumption 3 (Regularity and smoothness)

(a) Conditional on $g_i^0, \mu_{g_i^0}^0, \xi_i^0$, X_{i2} admits a uniformly continuous density function $f_{X_{i2}|g_i^0,\mu_{g_i^0}^0,\xi_i^0}$ such that $0 < \underline{\delta} \leq \inf_{x \in \mathcal{X}^0} f_{X_{i2}|g_i^0,\mu_{g_i^0}^0,\xi_i^0}(x) \leq \sup_{x \in \mathcal{X}^0} f_{X_{i2}|g_i^0,\mu_{g_i^0}^0,\xi_i^0}(x) \leq \overline{\delta} < \infty$.

(b) Almost surely, $\mathbb{E}\left(\left\|X_{12}\right\|^2 \mid g_1^0, \alpha^0, \lambda^0, \mu^0\right)$ is finite and $\mathbb{E}\left(X_{12}X_{12}' \mid g_1^0, \alpha^0, \lambda^0, \mu^0\right)$ is nonsingular.

(c) $\sum_{y \in \mathcal{Y}} y h^0(y, \cdot)$ is differentiable on $\mathbb{R}$ and not constant on the support of $X_{it}'\beta^0 + \alpha_{g_i^0 t}^0$.

Assumption 3 collects sufficient technical conditions that are useful to achieve point identification of β^0, α^0 given that h^0 is unknown, by relying on existing results in Ichimura (1993) for nonparametric i.i.d. single index models. In particular, it requires continuous covariates (which could be relaxed at the expense of heavier conditions) and invertibility of conditional Gram matrices.

Assumption 4 (Monotonicity) *There exists $y \in \mathcal{Y}$ such that $h^0(y, v)$ is strictly monotonic in v.*

Assumption 4 is a shape restriction which may be expected to hold at boundary points of $\mathcal{Y}$ (e.g., outside option in random utility models, absence of trade, absence of patenting in a count outcome model). Shape restrictions such as monotonicity have been routinely used to obtain point-identification in nonseparable panel data models.[23] This condition is weaker than log-concavity assumptions found in the literature (see, e.g. Chen, Fernández-Val, and Weidner, 2021; Bonhomme, Lamadon, and Manresa, 2022) that impose strongly unimodal densities (see Ibragimov, 1956).

[23] See, among many others, Klein and Spady (1993); Altonji and Matzkin (2005); Athey and Imbens (2006); Evdokimov (2011); Mugnier and Wang (2022).

Assumption 5 (Compensating variations) *For all fixed $(g, \widetilde{g}, t)$, there exist $x_1, x_2 \in \mathcal{X}$ such that*

$$\alpha^0_{\widetilde{g}t} + x_1'\beta^0 = \alpha^0_{gt} + x_2'\beta^0. \tag{8}$$

Similarly, for all $(g, t, \widetilde{t})$, there exist $x_3, x_4 \in \mathcal{X}$ such that

$$\alpha^0_{g\widetilde{t}} + x_3'\beta^0 = \alpha^0_{gt} + x_4'\beta^0. \tag{9}$$

Assumption 5 requires sufficient variation in the covariates and has the same flavor as the *compensating variations* used in D'Haultfoeuille, Hoderlein, and Sasaki (2021) and Mugnier and Wang (2022). As in the latter paper, it does not necessarily require a covariate with large support (it depends on the joint support of covariates and the unobserved group-specific effects), and ensures that there is overlap in the single index across unobserved clusters (not individuals) and periods. Theorem 1 below is the main identification result of the paper. Let $W^0_N = \left(\mathbf{1}\left\{g^0_i = g^0_j\right\}\right)_{(i,j)\in\{1,\ldots,N\}^2}$.

Theorem 1 *Let Assumptions 1-3(a) hold, and let N and T diverge jointly to infinity. Then,*

1. *$(W^0_N)_{N\in\mathbb{N}^*}$ and G^0 are point identified.*
2. *If Assumptions 3(b)-5 further hold, then h^0, β^0, and $(\alpha^0_{gt})_{(g,t)\in\mathcal{G}^0\times\mathbb{N}^*}$ are point identified.*

For the proof see Appendix A.1.

Remark 3 *A key argument of the proof of Theorem 1 is to frame the identification of the clustering γ^0 up to cluster relabeling as the equivalent problem of recovering the lower (or uppper)-triangular submatrix of the adjacency matrix W^0_N of the undirected graph $\mathcal{G}_N = \{V, E\}$ whose set of vertices V contains units $i \in \mathcal{N}$ and whose edges E contains all $(i, j) \in \mathcal{N}^2$ such that $g^0_i = g^0_j$. Given the clustering structure of the model, note that W^0_N has rank $R_N \leq G^0$ which is also its number of distinct rows because clusters form disconnected cliques in $\mathcal{G}_N$.*[24] *In other words, it is easily seen that identification of γ^0 up to cluster relabeling is equivalent to identification of all sets $\mathcal{C}^0(i) := \left\{j \in \mathcal{N} : g^0_j = g^0_i\right\}$ for $i \in \mathcal{N}$. Such a characterization has two advantages: (i) it is invariant to clusters relabeling and (ii) it reduces the NP-hard G^0-mean clustering problem to that of solving $N(N-1)/2$ pairwise binary classification problems.*[25] *Once the clustering γ^0 has been identified for all N, identification of G^0 follows easily by letting $N \to \infty$. Identification of β^0 can be obtained relying on within-cluster cross-sectional variation for a single cluster and time period and a result by Ichimura (1993) for a large class of cross-sectional nonparametric single-index models. Identification of cluster-specific effects and link function h^0 relies on the compensating variations and monotonicity of $h^0(y, \cdot)$ for some $y \in \mathcal{Y}$.*

[24] The related problem of "community detection" in networks has been widely studied in the statistical learning literature, and in particular in the compressed sensing literature. I do not pursue adaptation of spectral clustering techniques or recent development in Graph-cut problems for which very few asymptotic results in statistical settings with complex structure of dependencies are known. See von Luxburg (2007); Wang and Su (2021).

[25] Building on this insight, Mugnier (2022) proposes computationally straightforward pairwise differencing estimators for linear grouped fixed effects models. A similar-in-philosophy though different trick to break NP hardness is the binary segmentation approach of Wang and Su (2021).

A natural nonparametric estimation approach follows from the constructive identification strategy. Yet, it has the drawback of requiring a lot of nonparametric density estimation, i.e., a lot of tuning parameters as it requires combining nonparametric estimators for many unknown conditionals probabilities. This is similar to Gao, Li, and Xu (2022)'s approach in a pure network setting. I do not pursue the theoretical analysis of an estimator of this type, because I aim at developing a simple method for which inference tools are available. An open question is how the pairwise approach compares to the bruteforce fully nonparametric maximum likelihood approach. I note that, for a class of nonlinear (exponential) directed network models, the pairwise differencing approach developed in Mugnier (2022) yields a convenient estimation procedure under conditional moment restrictions, without requiring any nonparametric estimation, which reconciles computational simplicity and powerful inference.

4 Semiparametric Estimation

In the first part of this section, I propose a general M-estimation framework accomodating nonlinear models when the number of clusters, $G^0 \in \mathbb{N}^*$, is known to the researcher.[26] In the second part, I specialize the framework to cases where $h^0 \in \mathcal{H}$ is further assumed to be known (e.g., Probit, Logit, Poisson) to define semiparametric NGFE estimators. In the third part, I discuss computation.

4.1 A Generic M-Estimation Framework

Assume from now that $G^0 \in \mathbb{N}^*$ is known to the researcher, and suppose there exists a known function $\rho : \mathcal{Y} \times \mathcal{X} \times \mathcal{B} \times \mathcal{H} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T} \longrightarrow \mathbb{R}$ such that $\theta^0_{NT} := (\beta^{0\prime}, h^0, \gamma^{0\prime}, \alpha^{0\prime})'$ satisfies

$$\theta^0_{NT} = \underset{\theta \in \mathcal{B} \times \mathcal{H} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T}}{\arg\max} \; \mathbb{E}\left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho(Y_{it}, X_{it}; \theta) \mid \gamma, \alpha \right), \tag{10}$$

where $\mathcal{G}^{0N} = \{1, ..., G^0\}^N$ is the set of all partitions of $\{1, ..., N\}$ into at most G^0 clusters. Provided it exists, the M-NGFE nonparametric estimator $\widehat{\theta}^{\mathrm{M}}_{\rho} := (\widehat{\beta}^{\mathrm{M}\prime}, \widehat{h}^{\mathrm{M}}, \widehat{\gamma}^{\mathrm{M}\prime}, \widehat{\alpha}^{\mathrm{M}\prime})'$ of θ^0_{NT} solves

$$\widehat{\theta}^{\mathrm{M}}_{\rho} \in \underset{\theta \in \mathcal{B} \times \mathcal{H} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T}}{\arg\max} \; \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho(Y_{it}, X_{it}; \theta). \tag{11}$$

Finding a suitable ρ-function, proving identification of θ^0_{NT} (i.e., that eq. (10) holds), and deriving the asymptotic properties of the sequence of $\widehat{\theta}^{\mathrm{M}}_{\rho}$ are certainly difficult problems beyond the scope

[26] Estimating G^0 in nonlinear models with time-varying unobserved heterogeneity is a difficult problem that is beyond the scope of the paper. See Chen, Fernández-Val, and Weidner (2021) for a discussion in some concave nonlinear factor type models. An AIC or BIC-type criterion à la Bonhomme and Manresa (2015); Bai and Ng (2002) could be employed but would require to know at least an upper bound on G^0. Letting G^0 grow slowly with N, T could also be of interest but would require a different analysis that is beyond the scope of the paper. Note that Bonhomme, Lamadon, and Manresa (2022) *need* the number of clusters to increase as they assume a (possibly) continuous underlying unobserved heterogeneity.

of the paper, each of them would require further assumptions. Moreover, computation of $\widehat{\theta}_\rho^{\mathrm{M}}$ is generally infeasible because maximization problem (11) is a non-smooth non-concave optimization problem with combinatorial optimization (due to the clustering part) over an infinite-dimensional space (due to $\mathcal{H}$). A practical solution to make the problem finite-dimensional is sieve-estimation of h^0 but this is beyond the scope of this paper. Instead, I focus on semiparametric versions where h^0 is assumed to be known and that are of practical interest in many empirical applications.

4.2 Semiparametric NGFE Estimators

From now on, I assume that $h^0 \in \mathcal{H}$ is known (e.g., Logit, Probit, Poisson, etc.) and consider the problem of estimating $\theta^0_{NT} := (\beta^{0\prime}, \gamma^{0\prime}, \alpha^{0\prime})'$ in the semiparametric model (2) with known G^0. The semiparametric NGFE estimator of θ^0_{NT}, denoted $\widehat{\theta}^{\mathrm{NGFE}} := (\widehat{\theta}', \widehat{\gamma}', \widehat{\alpha}')'$, is the M-NGFE estimator $\widehat{\theta}_\rho^{\mathrm{M}}$ (once suppressing dependence on h) with $\rho(Y_{it}, X_{it}; \theta) = \ln h^0\left(Y_{it}, X_{it}'\beta + \alpha_{g_i t}\right)$. In other words, $\widehat{\theta}^{\mathrm{NGFE}}$ is solution to the following minimization problem:

$$\widehat{\theta}^{\mathrm{NGFE}} \in \underset{\theta \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T}}{\arg\min} \; \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} - \ln h^0\left(Y_{it}, X_{it}'\beta + \alpha_{g_i t}\right), \tag{12}$$

where the minimum is taken over all possible common parameters β, partitions $\gamma = (g_1, ..., g_N)'$ of the N individuals into G^0 clusters, and cluster-specific time effects $\{\alpha_{gt} : (g, t)\}$. Note that the NGFE estimator is a "classification likelihood" estimator. For given values of β and α, the optimal cluster assignment for individual i is

$$\widehat{g}_i(\beta, \alpha) = \underset{g \in \mathcal{G}^0}{\arg\min} \; \frac{1}{NT} \sum_{t=1}^{N} \sum_{t=1}^{T} - \ln h^0\left(Y_{it}, X_{it}'\beta + \alpha_{gt}\right), \tag{13}$$

where the minimum g is taken in case of a non-unique solution. The NGFE estimator of $(\beta^{0\prime}, \alpha^{0\prime})'$ in (12) can then be written as

$$(\widehat{\beta}, \widehat{\alpha}) = \underset{(\beta, \alpha) \in \mathcal{B} \times \mathcal{A}^{G^0 T}}{\arg\min} \; \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} - \ln h^0\left(Y_{it}, X_{it}'\beta + \alpha_{\widehat{g}_i(\beta,\alpha) t}\right), \tag{14}$$

where $\widehat{g}_i(\beta, \alpha)$ is given by (13).

4.3 Computation

The minimization problem (12) is not differentiable nor convex in θ. In particular, it may be subject to the existence of local minima. Note that the number of partitions of N individuals into G^0 clusters increases steeply with N, making exhaustive search impossible.[27] I propose the following simple

[27] The number of partitions of N objects into G^0 disjoint and non-empty subsets is $\frac{1}{N!}\sum_{i=1}^{N}(-1)^{N-i}\binom{N}{i}N^{G^0} \propto \frac{G^{0\,N}}{G^0!}$. In fact the G^0-means problem without regressors in a cross-sectional setting is NP-hard (see, e.g., Aloise, Deshpande, Hansen, and Popat, 2009).

algorithm which is an extension of the popular Lloyd (1982)'s algorithm for k-means, a "greedy" algorithm providing a converging sequence of heuristic solutions in polynomial time.

Iterative Algorithm:

1. Let $(\beta^{(0)}, \alpha^{(0)}) \in \mathcal{B} \times \mathcal{A}^{G^0 T}$ be some starting value. Set $s = 0$.

2. Compute for all $i \in \{1, ..., N\}$:

$$g_i^{(s+1)} = \underset{g \in \mathcal{G}^0}{\arg\min} \sum_{t=1}^{T} - \ln h^0 \left(Y_{it}, X_{it}' \beta^{(s)} + \alpha_{gt}^{(s)} \right). \tag{15}$$

3. Compute:

$$\left(\beta^{(s+1)}, \alpha^{(s+1)} \right) = \underset{(\beta, \alpha) \in \mathcal{B} \times \mathcal{A}^{G^0 T}}{\arg\min} \sum_{i=1}^{N} \sum_{t=1}^{T} - \ln h^0 \left(Y_{it}, X_{it}' \beta + \alpha_{g_i^{(s+1)} t} \right). \tag{16}$$

4. Set $s = s + 1$ and go to Step 2 (until numerical convergence).

Algorithm 1 alternates between two steps. In the "assignment" step, each individual i is assigned to cluster g_i whose vector of time effects minimizes individual's i time-averaged log-likelihood given the slope parameter. In the "update step", β and α are computed using maximum likelihood and controlling for interactions of cluster and time dummies. A potential issue is that the solution depends on the chosen starting values. Drawing starting values at random and selecting the solution that yields the lowest objective is a practical solution in low-dimensional problems. Finding a fast approximation of NGFE for larger-scale problems and controlling its optimization error is left for further research.[28]

5 Asymptotic Properties of Semiparametric NGFE Estimators

In this section, I assume that $\theta_{NT}^0 := (\beta^{0\prime}, \alpha^{0\prime}, \gamma^{0\prime})'$ is identified (e.g., by Theorem 1) and derive the asymptotic properties of semiparametric NGFE estimators. I consider an asymptotic framework where N and T tend jointly to infinity but G^0 does not grow with N and T. I focus on binary choice models with grouped fixed effects as the leading case. Similar results can be obtained for other models with strictly concave log-likelihood function (see Appendix B.4), but I stick to binary choice models to keep the exposition simple. I abstract from optimization errors and study the asymptotic behaviour of the exact sequence of estimates defined in eq. (12).

[28] Note that an algorithm similar to Algorithm 2 in Bonhomme and Manresa (2015) can be employed to improve the trade-off between exploration and exploitation during the optimization process.

5.1 Binary Choice Model With Grouped Fixed Effects

Consider the following data generating process:

$$Y_{it} = \mathbf{1}\left\{X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} \geq 0\right\}, \quad i = 1, ..., N,\ t = 1, ..., T. \tag{17}$$

For any $\mathbf{Z} := \{Z_{it} : (i,t)\}$, let $\mathbf{Z}_-^{(t)} = \{Z_{is} : 1 \leq i \leq N, 1 \leq s \leq t\}$ and $\mathbf{Z}_+^{(t)} = \{Z_{is} : 1 \leq i \leq N, t \leq s \leq T\}$.

Assumption 6
Eq. (17) holds and

(a) For all t: $\left(\mathbf{X}_-^{(t)}, \gamma^0, \alpha^0, \boldsymbol{\varepsilon}_-^{(t-1)}\right)$ and $\boldsymbol{\varepsilon}_+^{(t)}$ are independent.[29]

(b) The $\{\varepsilon_{it} : (i,t)\}$ are identically distributed with known cumulative distribution function Ψ that is fully supported on $\mathbb{R}$, twicely continuously differentiable, strictly increasing, and such that $(\ln \Psi)'' < 0$. Moreover, Ψ' is symmetric around 0.

Assumption 6(a) is a weak exogeneity assumption, standard in the panel data literature, which allows X_{it} to contain predetermined variables with respect to Y_{it}. In particular, X_{it} can include lags of Y_{it} to accommodate dynamic models. This assumption does not restrict the correlation between (γ^0, α^0) and $\{\mathbf{X}_i : i\}$. Assumption 6(b) is standard in semiparametric panel discrete choice models and yields strict concavity of the log-likelihood function under minimal amount of cluster-specific and time-specific variation in the covariates (as assumed, e.g., in Fernández-Val and Weidner, 2016; Bonhomme, Lamadon, and Manresa, 2022; Chen, Fernández-Val, and Weidner, 2021).[30] The second part of Assumption 6(b) is weak and is statisfied by the Probit ($\Psi(u) = \int_{-\infty}^u (1/\sqrt{2\pi})e^{-t^2/2}dt$) and Logit ($\Psi(u) = 1/(1+e^{-u})$) distributions. Symmetry of Ψ is not necessary but it conveniently simplifies notation in the proofs. Under Assumption 6, note that eq. (17) is a semiparametric NGFE model (2) with known link function $h^0(y,z) = \Psi(z)^{\mathbf{1}\{y=1\}}(1-\Psi(z))^{\mathbf{1}\{y=0\}}$. The corresponding NGFE estimator writes

$$(\widehat{\beta}, \widehat{\gamma}, \widehat{\alpha}) \in \underset{(\beta,\gamma,\alpha)\in\mathcal{B}\times\mathcal{G}^{0N}\times\mathcal{A}^{G^0T}}{\arg\min} \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T -\ln\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{g_i t}\right)\right), \tag{18}$$

where $Q_{it} = 2Y_{it} - 1$.

5.2 Consistency

Consider the following assumption.

Assumption 7

(a) $\mathcal{B}$ and $\mathcal{A}$ are compact convex subsets of $\mathbb{R}^p$ and $\mathbb{R}$, respectively.

---

[29] If one lag Y_{it-1} is included as regressor, I assume that Y_{i0} is observed and contained in $\mathbf{X}_-^{(t)}$. Higher-order dependence can be accommodated similarly.

[30] See also, Pratt (1981).

(b) *There exists a constant $M > 0$ such that* $\left\|X_{it}\right\| \leq M$ *almost surely.*

(c) *Let $\overline{X}_{g \wedge \widetilde{g},t}$ denotes the mean of X_{it} in the intersection of clusters $g_i^0 = g$, and $g_i = \widetilde{g}$. For all partitions $\gamma = \{g_1, ..., g_N\} \in \Gamma_{\mathcal{G}^{0N}}$, let $\widehat{\rho}(\gamma)$ denote the minimum eigenvalue of the following matrix:*

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it} - \overline{X}_{g_i^0 \wedge g_i,t})(X_{it} - \overline{X}_{g_i^0 \wedge g_i,t})'.$$

Then, $\text{plim}_{N,T\to\infty} \min_{\gamma \in \Gamma_{G^0}} \widehat{\rho}(\gamma) = \rho > 0$.

Assumption 7(a) is standard in the context of M-estimation. Assumption 7(b) is for a matter of convenience (it simplifies the proof). It strengthens Assumption 1(b) in Bonhomme and Manresa (2015), and ensures (together with Assumption 7(a)) strong concavity of the log-likelihood function and rules non-stationary covariates.[31] Assumption 7(c) is the same noncollinearity condition as Assumption 1(g) in Bonhomme and Manresa (2015). It requires that X_{it} shows sufficient within-cluster variation over time and across individuals, and is related to standard noncolinearity assumptions encountered in the large-N, large-T panel data literature (see, e.g., Bai, 2009; Chen, Fernández-Val, and Weidner, 2021; Vogt and Linton, 2017; Ando and Bai, 2022). It allows for time-invariant covariates provided that they have a sufficiently rich support. As a special case highlighted in Bonhomme and Manresa (2015), Assumption 7(c) is satisfied if X_{it} are discrete and, for all g, the conditional distribution of X_i given $g_i^0 = g$ has strictly more than G^0 points of supports.

Theorem 2 (Consistency) *Let Assumptions 6 and 7 hold. Then, as N and T tend to infinity:*

1. $\widehat{\beta} \xrightarrow{p} \beta^0$.
2. $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2 \xrightarrow{p} 0$.

For the proof see Appendix A.3.

Theorem 2 shows that NGFE estimators of the common slope coefficient and cluster-specific effects in NGFE binary choice models are both consistent.

5.3 Asymptotic Distribution

Consider the following assumption.

Assumption 8

(a) *For all $g \in \mathcal{G}^0$:* $\text{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} = \pi_g > 0$.

(b) *For all $(g, \widetilde{g}) \in \mathcal{G}^{02}$ such that $g \neq \widetilde{g}$:* $\text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})^2 = c_{g,\widetilde{g}} > 0$.

[31] One could relax this assumption by allowing covariates to have sub-gaussian tails (see, e.g., Vershynin, 2019, for a definition). I do not pursue this avenue in order to keep the exposition light. Moment conditions in Bonhomme and Manresa (2015) also rule out non-stationary covariates.

(c) *There exist constants $a > 0$ and $d > 0$ and a sequence $\alpha(t) \leq \exp(-at^d)$ such that, for all $i \in \{1, ..., N\}$ and $(g, \widetilde{g}) \in \mathcal{G}^{02}$ such that $g \neq \widetilde{g}$, $\left\{\alpha^0_{gt} - \alpha^0_{\widetilde{g}t} : t\right\}$ are strongly mixing processes with mixing coefficients $\alpha(t)$.*

Assumptions 8(a)-(c) are identical to Assumptions 2(a)-(c) in Bonhomme and Manresa (2015), respectively. Assumption 8(a) ensures that no cluster is asymptotically negligible relative to the others and that each cluster has a large number of observations in the population. This is equivalent to the "strong factor" condition in approximate factor models (see, e.g., Assumption 1.(v) in Chen, Fernández-Val, and Weidner, 2021). Assumption 8(b) imposes that the G^0 clusters are well separated in the population. As discussed in a recent work by Chetverikov and Manresa (2021), departing from such an assumption seems quite difficult. Assumption 8(c) restricts the dependence and tail properties of the processes $(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})$, which are assumed to be strongly mixing. Assumption 8(d) is standard and requires a sufficient amount of variation in the covariates.

Assumption 8 allows me to rely on exponential inequalities for dependent processes (e.g., Rio, 2000) in order to bound misclassification probabilities, almost the same way as in the proof of Theorem 2 in Bonhomme and Manresa (2015). The novelty here is that their assumption that the idiosyncratic shock in the linear model is a strong mixing process is hidden in the parametric and independence restrictions formulated in Assumption 6, the latter allowing to rely on exponential inequalities for martingale differences (see, e.g., Lesigne and Volný, 2001) to control remainder terms in the proofs (essentially the score).

Let $(\widetilde{\beta}, \widetilde{\alpha})$ be such an infeasible version of the NGFE estimator where cluster membership g_i, instead of being estimated, is fixed to its population counterpart g^0_i:

$$(\widetilde{\beta}, \widetilde{\alpha}) = \underset{(\beta,\alpha)\in\mathcal{B}\times\mathcal{A}^{G^0T}}{\operatorname{argmin}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} - \ln \Psi\left(Q_{it}\left(X'_{it}\beta + \alpha_{g^0_i t}\right)\right). \tag{19}$$

This is the maximum likelihood estimator in the pooled regression of Y_{it} on X_{it} and the interactions of population cluster dummies and time dummies.

Assumptions 6, 7, and 8 provide conditions under which estimated cluster memberships converge to their population counterparts, and the NGFE estimator defined in (18) is asymptotically equivalent to the infeasible maximum likelihood estimator $(\widetilde{\beta}, \widetilde{\alpha})$, when N and T tend to infinity and $N/T^\nu \to 0$ for some $\nu > 0$ (see Lemma 7 in Appendix A.4.1). In particular, this allows T to grow considerably more slowly than N. Because of invariance to relabeling of the clusters, the results for cluster membership and cluster-specific effects are understood to hold given a suitable choice of the labels (see the proof for details). Theorem 2 and eq. (53) crucially hinge on the restrictive assumption that the number of well-separated clusters G^0 is known and fixed, but it could be that consistent estimation of $\widehat{\beta}$ remains possible under weaker assumptions that would nonetheless prevent consistent estimation of cluster memberships.[32]

Given Lemma 7, showing asymptotic normality of the NGFE estimator then reduces to the

[32] I thank Martin Weidner for pointing out this to me, something also discussed in Dzemski and Okui (2018).

simpler problem of showing asymptotic normality of the infeasible (oracle) MLE $(\widetilde{\beta}, \widetilde{\alpha})$. Let $Z_{it}^0 = X_{it}'\beta^0 + \alpha_{g_i^0 t}^0$. For all $g \in \mathcal{G}$, all $t \in \{1, ..., T\}$, let $\widetilde{X}_{gt}$ denote the projection of X_{it} on the space spanned by the cluster membership variable under a metric weighted by $(-\ln \Psi)''(Q_{it}Z_{it}^0)$:

$$\widetilde{X}_{gt} = \left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}(\ln \Psi)''(Q_{it}Z_{it}^0)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}(\ln \Psi)''(Q_{it}Z_{it}^0)X_{it}\right),$$

i.e., the weighted average of X_{it} for individuals $\{i : g_i^0 = g\}$. Also, let $\widehat{\pi}_{gt}$ denote the following weighted average:

$$\widehat{\pi}_{gt} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}(-\ln \Psi)''(Q_{it}Z_{it}^0).$$

Assumption 9 below allows to characterize the asymptotic distribution of the infeasible MLE $(\widetilde{\beta}, \widetilde{\alpha})$.

Assumption 9

(a) $\{Y_{it} : (i,t)\}$ *are independent conditional on* $(\mathbf{X}, \gamma^0, \alpha^0)$.

(b) There exists a positive definite matrix Σ_β *such that*

$$\Sigma_\beta = \text{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T(-\ln \Psi)''(Q_{it}Z_{it}^0)\left[X_{it} - \widetilde{X}_{g_i^0 t}\right]\left[X_{it} - \widetilde{X}_{g_i^0 t}\right]'.$$

(c) As N *and* T *tend to infinity,*

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T\left\{(-\ln \Psi)''(Q_{it}Z_{it}^0)\left(X_{it} - \widetilde{X}_{g_i^0 t}\right)\right\}\left\{Q_{it}(-\ln \Psi)'(Q_{it}Z_{it}^0)\right\} \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta).$$

(d) For all (g,t)*:* $\text{plim}_{N\to\infty}\widehat{\pi}_{gt} = \widetilde{\pi}_{gt} > 0$.

(e) For all (g,t)*:*

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \mathbb{E}\left(\mathbf{1}\left\{g_i^0 = g\right\}\mathbf{1}\left\{g_j^0 = g\right\}Q_{it}Q_{jt}(\ln \Psi)'(Q_{it}Z_{it}^0)(\ln \Psi)'(Q_{jt}Z_{jt}^0)\right) = \omega_{gt} > 0.$$

(f) For all (g,t)*, and as* N *and* T *tend to infinity:*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}Q_{it}(\ln \Psi)'(Q_{it}Z_{it}^0) \xrightarrow{d} \mathcal{N}(0, \omega_{gt}).$$

(g) The true value of β*,* β^0*, is in the interior of* $\mathcal{B}$*. For all* T*, the true value of* α*,* α^0*, is in the interior of* $\mathcal{A}^{G^0 T}$*.*

Assumption 9(a) rules out dynamic or feedbacks.

Theorem 3 (Asymptotic Distribution) *Let Assumptions 6-9 hold and let N and T tend to infinity such that $N/T \to \infty$ and, for some $\nu > 1$, $N/T^\nu \to 0$. Then:*

$$\sqrt{NT}(\widehat{\beta} - \beta^0) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta^{-1}\right), \tag{20}$$

and, for all (g,t),

$$\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha^0_{gt}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\widetilde{\pi}^2_{gt}}\right), \tag{21}$$

where Σ_β, ω_{gt}, and $\widetilde{\pi}_{gt}$ are defined in Assumption 9.

For the proof see Appendix A.4.2.

Theorem 3 demonstrates that NGFE estimators in NGFE binary choice models achieve the parametric $\sqrt{NT}$ and $\sqrt{N}$ rates of convergence and are free of Neyman and Scott (1948)'s incidental parameters problem. The asymptotic regime $N/T \to 0$ is needed since (i) there are time effects and (ii) the model is nonlinear. These rates are in contrast with standard interactive fixed-effects models (see, e.g. Bai, 2003, 2009; Ando and Bai, 2022) for which $\sqrt{N}$ consistency of the time-varying factors requires $N/T^2 \to 0$ or more generally $N/T \to \kappa$, $0 < \kappa < \infty$, as it is assumed for instance in Fernández-Val and Weidner (2016); Chen, Fernández-Val, and Weidner (2021). The intuition behind this result is that the extreme sparsity of the factor loading structure in model (17) allows NGFE estimators to achieve fast accurate classification of individuals, which reduces the estimation problem to that of a standard nonlinear models with multidimensional time-varying fixed effect in the limit.[33] Consistent estimators of the asymptotic variances are given in Appendix C.

5.4 Average Partial Effects (APEs)

Under Assumption 6, if $X_{it,k}$, the kth element of X_{it} is binary, its partial effect on the conditional probability of Y_{it} is

$$\Delta(X_{it}, \beta^0, \alpha^0_{g_i^0 t}) = \Psi(\beta^0_k + X'_{it,-k}\beta^0_{-k} + \alpha^0_{g_i^0 t}) - \Psi(X'_{it,-k}\beta^0_{-k} + \alpha^0_{g_i^0 t}),$$

where β^0_k is the kth element of β^0, and $X_{it,-k}$ and β^0_{-k} include all elements of X_{it} and β^0 except the kth element. If $X_{it,k}$ is continuous, the partial effect of $X_{it,k}$ on the conditional probability of Y_{it} is

$$\Delta(X_{it}, \beta^0, \alpha^0_{g_i^0 t}) = \beta^0_k \Psi'(X'_{it}\beta^0 + \alpha_{g_i^0 t}),$$

where Ψ' is the derivative of Ψ. As discussed in Fernández-Val and Weidner (2016), if $(X_{it}, g_i^0, (\alpha^0_{gt})_{g \in \mathcal{G}^0})$ is identically distributed over i but can be heterogeneously distributed over t, then $\mathbb{E}(\Delta_{it}) = \delta^0_t$ and

[33] To see the factor-loading structure, note that model (17) can be written as $Y_{it} = \mathbf{1}\{X'_{it}\beta + \lambda'_i f_t - \varepsilon_{it} \geq 0\}$, where $\lambda'_i = (\mathbf{1}\{g_i^0 = 1\}, ..., \mathbf{1}\{g_i^0 = G^0\}) \in \left\{b \in \{0,1\}^{G^0} : \sum_{g=1}^{G^0} b_g = 1\right\}$ and $f_t = (\alpha^0_{gt})'_{g \in \mathcal{G}^0} \in \mathcal{A}^{G^0}$. If $N/T \to \kappa \in (0, +\infty)$, similar arguments than Chen, Fernández-Val, and Weidner (2021) apply and bias-correction methods are needed.

$\delta^0_{NT} = \frac{1}{T}\sum_{t=1}^{T}\delta^0_t$ changes only with T. If $(X_{it}, g_i^0, (\alpha^0_{gt})_{g\in\mathcal{G}^0})$ is identically distributed over i and stationary over t, then $\mathbb{E}(\Delta_{it}) = \delta^0_{NT}$, and $\delta^0_{NT} = \delta^0$ does not change with N and T.

Deriving the asymptotic properties of plug-in estimators of average partial effects of the type $\widehat{\delta}_{NT} = \Delta(\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma})$ should follow similar arguments as in Fernández-Val and Weidner (2016).

6 Monte Carlo Simulations

In this section, I conduct Monte Carlo experiments to assess the numerical performance of NGFE estimators in finite samples, in terms of bias, root mean squared errors (RMSE), classification (Precision, Recall, Rand Index), execution (CPU) time, and inference accuracy (standard errors, standard deviation and coverage). I compare the results with currently available competitors. I consider Chamberlain (1980); Rasch (1960)'s conditional logit (CMLE), nonlinear two-way fixed effects (NLTWFE, see, e.g. Fernández-Val and Weidner, 2016; Mugnier and Wang, 2022), Bonhomme, Lamadon, and Manresa (2022)'s 2-step grouped fixed effects (2GFE), pooled OLS regression, linear two-way fixed effects (LTWFE), and Bonhomme and Manresa (2015)'s GFE estimators.[34]

As in Bonhomme and Manresa (2015), I focus on settings of moderate size ($N = 90$, $T = 7$) to highlight the performance of NGFE with small datasets as often encountered in macro/meso-economics (e.g., in my empirical application). Having large N is not computationally demanding. When T is very large, computation of the NGFE estimate might be demanding and results in Mugnier (2022) could probably be adapted. I consider static and dynamic logit models, and four DGPs for the time-varying covariates (more or less correlated with the unobserved heterogeneity, UH hereafter), where the number of groups G^0 each time varies across $\{2, 3, 5\}$. Variation across time periods in the covariates is not necessary for NGFE but allows for comparisons (e.g., with CMLE).

Overall, I find that NGFE estimators perform best uniformly across competitors in the design they are meant to adress: correlated time-varying unobserved heterogeneity (DGP 1). In other DGPs, where the unobserved heterogeneity does not vary with time, they might be slightly more noisy than well-suited estimators (e.g., CMLE or NLTWFE) and have a larger finite sample bias.

6.1 Static Logit Model

The data generating process is

$$Y_{it} = \mathbf{1}\left\{X_{it}\beta + \alpha_{g_i t} > \varepsilon_{it}\right\}, \quad i = 1, ..., N,\ t = 1, ..., T, \tag{22}$$

where $\beta = 1$ and ε_{it} ~Logit$(0, \pi^2/3)$, g_i ~Unif$\{1, ..., G^0\}$ for $G^0 \in \{2, 3, 5\}$, and, letting with $\mu = (-1, 1)'$ if $G^0 = 2$, $\mu = (-\pi/\sqrt{3}, 0, \pi/\sqrt{3})'$ if $G^0 = 3$, and $\mu = (-2\pi/\sqrt{3}, -\pi/\sqrt{3}, 0, \pi/\sqrt{3}, 2\pi/\sqrt{3})'$

[34] I leave comparison with Charbonneau (2017)'s conditional logit and Chen, Fernández-Val, and Weidner (2021)'s nonlinear factor models for further research. A definition of the metrics and more details are given in Appendix D.

if $G^0 = 5$, V_i such that $\Pr(V_i = -2) = 1/12, \Pr(V_i = -1) = 1/4, \Pr(V_i = 0) = 1/3, \Pr(V_i = 1) = 1/4, \Pr(V_i = 2) = 1/12$, and $W_{it} \sim \mathcal{N}(0, 1)$:

- DGP1 (grouped patterns of time-varying UH): $\alpha_{g0} = \mu_g$, for $t \geq 1$, $\alpha_{gt} = 0.1\alpha_{gt-1} + (-1)^{g-1}U_{gt}$, $U_{gt} \sim$Unif$[0, 1]$, $X_{it} = 0.5V_i + 0.8U_{g_i^0 t}$.
- DGP2 (grouped patterns of time-invariant UH): $\alpha_{gt} = \mu_g$, $X_{it} = 0.3\mu_{g_i} + V_i + 0.8W_{it}$.
- DGP3 (continuous time-invariant UH): $\alpha_i \sim \mathcal{N}(0, 1)$, $X_{it} = \alpha_i + 0.5V_i + 0.8W_{it}$.
- DGP4 (No UH): $\alpha_{gt} = 0$, $X_{it} = 0.5V_i + 0.8W_{it}$.

The variables U_{gt}, V_i, W_{it}, g_i and ε_{it} are independent and i.i.d. across individuals and time periods. All the results are based on 50 Monte-Carlo replications and computed using Algorithm 1 with 200 randomized initialization points (results improve by increasing this number).

Table 1 reports the bias and RMSE of NGFE and five competing estimators. It shows that NGFE estimates minimize both metrics across all estimators in DGP 1 (e.g., one order of magnitude less than CMLE or 2STEPGFE, the best competitors). If there is no UH (DGP 4), NGFE keeps a reasonable RMSE compared to CMLE but has small bias (e.g. RMSE of .151 v.s. .152 if $G^0 = 2$ and .178 v.s. .118 if $G^0 = 5$, Bias of 0.040 v.s. -0.002 and 0.114 v.s. 0.018 respectively). All linear estimators perform very poorly. The 2-step GFE is more noisy in general.

Table 2 shows that any measure of the clustering accuracy remains at a high level because of the high level of UH. For instance, the misclassification rate falls below 50% when $G^0 = 2$ only. Unreported simulations show that it actually drops to 5% when $G^0 = 2$ and cluster-specific effects are not correlated with the covariates. There is a continum between the two regimes that merits further investigation. Precision also improves with the number of iterations of Lloyd (1982)'s algorithm. The CPU time of the method is comparable to that of other clustering methods such as Bonhomme, Lamadon, and Manresa (2022)'s 2-step GFE.

Table 3 suggests that estimates of the standard errors based on the large-T clustered variance formula match on average the effective finite sample dispersion of the NGFE estimates. The resulting confidence intervals have an almost correct coverage though showing a small finite-sample under-coverage.[35] In particular, Table 3 suggests good coverage rates around the prescribed theoretical level of 95% (e.g., .86, .80, .84 in DGP 1 and .92, .92, .88 in DGP 4), which fall with the number of groups and, more generally, with the degree of continuity of the UH (e.g., below .5 in DGP 3 but still .82 in DGP 2 with $G^0 = 2$).

[35] A similar finite-sample undercoverage phenomenon is also reported in Bonhomme and Manresa (2015), who suggest the use of a bootstrap estimator instead.

6.2 Dynamic Logit Model

The data generating process is

$$\begin{aligned} Y_{it} &= \mathbf{1}\left\{Y_{it-1}\beta_1 + X_{it}\beta_2 + \alpha_{g_it} > \varepsilon_{it}\right\}, \quad i=1,...,N,\ t=1,...,T, \\ Y_{i0} &= \mathbf{1}\left\{X_{i0}\beta_2 + \alpha_{g_i0} > \varepsilon_{i0}\right\}, \quad i=1,...,N, \end{aligned} \tag{23}$$

where $\beta_1 = 0.5$ and $\beta_2 = 1$. Tables 4-6 report the same statistics as Tables 1-3 but for the dynamic model. Results for β_2 are very similar to that for β. On the other hand, the precision of NGFE estimates of β_1 is more mixed (the conditional independence assumption 9(a) does not hold here). Previous comparisons still apply there.

7 Empirical Application: Revisiting the Inverted-U Relationship Between Innovation and Competition

Does more competition lead to more innovation? This fundamental question (e.g., for Antitrust and Competition policy) has been the subject of a longstanding academic debate in the fields of industrial organization and macroeconomics of endogeneous growth theory (for surveys, see, e.g., Griffith and Van Reenen, 2021; Gilbert, 2006).[36] On the one hand, more competition reduces profit and postinnovation rents, and therefore disincentivizes innovation: this is the so-called *Schumpetarian effect*. On the other hand, more competition may reduce a firm's preinnovation rent by more than it reduces its postinnovation rent and thus foster innovation and growth: this is the *escape-competition effect*.

In an influential paper, Aghion, Bloom, Blundell, Griffith, and Howitt (2005)[ABBGH henceforth] reconcile these two contradictory views by documenting an inverted-U relationship between the number of citation-weighted patents and a measure of product market competition using a panel data set of seventeen UK industries (i) observed over the period 1973-1994 (t). The inverted-U shape is predicted by a model of endogeneous growth and estimated after controlling for multiplicatively separable industry and year fixed effects, aimed at capturing permanent unobserved technological levels and common trends. The authors' preferred specification is a conditional fixed effects (FE) Poisson model: for all $p \in \{0, 1, ...\}$

$$\begin{aligned} &\Pr(\text{cwpatent}_{it} = p \mid \text{comp}_{it}, \nu_i, \xi_t) \\ &\quad = \frac{\exp(p(g(\text{comp}_{it}) + \nu_i + \xi_t)\exp(-\exp(g(\text{comp}_{it}) + \nu_i + \xi_t)))}{p!}, \end{aligned} \tag{24}$$

where cwpatent_{it} represents the number of citation-weighted patents in industry i in year t, comp_{it}

[36] For public coverage, see, e.g, Lohr, Steeve "How Software Is Stifling Competition and Slowing Innovation", *The New York Times*, 7 Jul, 2022. Last consulted on September 29, 2022 at: https://www.nytimes.com/2022/07/21/business/software-james-bessen-book.html.

is one minus the average Lerner index in industry i in year t, ν_i is an unobserved industry-specific permanent level of innovation, ξ_t captures macroeconomic trend, and $g(\cdot)$ is a second-degree polynomial.[37] Figure 1 shows ABBGH's original inverted-U relationship, by replicating ABBGH's Figure II, a scatterplot comparing the fit of the exponential model (24) with that of a nonparametric spline.[38]

While model (24) is in line with a large body of the previous literature (see, e.g., Hausman, Hall, and Griliches, 1984; Gourieroux, Monfort, and Trognon, 1984), it imposes strong assumptions on the data generating process: conditional Poisson distribution and multiplicative separability of unobserved effects. In particular, the inverted-U relationship seems fragile as recent empirical research has reported both increasing and decreasing monotonic relationships depending on the controls included (Aghion, Van Reenen, and Zingales, 2013), whether accounting or not for the presence of structural breaks (Correa, 2012), or the country data used (Hashmi, 2013; Askenazy, Cahn, and Irac, 2013), etc. This has spurred a variety of explanations and theoretical models.

To the best of our knowledge, however, no paper has assessed the robustness of the inverted-U relationship to modeling choices regarding unobserved heterogeneity. As ABBGH and Correa (2012) argue, innovation is a dynamic process and endogeneity issues might comes from unobserved forces that drive both innovation and the market structure in a dynamic way.[39] Moreover, while industry might be a good level to control for permanent scaling, it is likely that among the 311 firms of the panel, a few time-varying paths emerge. A natural question is then: to which extent are all industries subject to the same economic trend (i.e., time effect) during the 1973-1994 period where, e.g., the development of I.T. has been exponential and plausibly shaped market structures?

In this section, I illustrate how the class of NGFE models together with semiparametric NGFE estimators introduced in this paper can be used to adress this question, challenging the view that firms are all subjects to the same macroeconomic trends and that the unobserved propensity to innovate and compete is industry-specific and fixed across time.

Data. I use ABBGH's original data set available at N. Bloom's website.[40] This is an unbalanced industry-level panel based on 311 firms listed on the London Stock Exchange and grouped in 17 two-digit SIC code industries, which received patent grants from the United States Patent and Trademark Office (USPTO). The period covered by the dataset is from 1973 until 1994 and there are 354 observations. In particular, here $N = 17$ and $T = 22$ and I assume that missing observations

[37] The fact that the number of patents is weighted and averaged at the industry level makes it a "continuous" variable with a mass point at 0. This is probably a reason why the authors apply a discrete model. See the summary statistics in Table 7. See Aghion, Bloom, Blundell, Griffith, and Howitt (2005) for details on the construction of each variable.

[38] I note that the scale of the y-axis in ABBGH's Figure II is incorrect, as well as the legend of their Figure I since the graph in fact corresponds to specification (1) in their Table I (and not (2) as claimed).

[39] Fernández-Val and Weidner (2016) estimate model (24), including one lag of the dependent variable as an additional regressor and find ABBGH's results to be robust to this change. Yet, unobserved time-varying heterogeneity could still remain.

[40] https://nbloom.people.stanford.edu/sites/g/files/sbiybj4746/f/abbgh.zip.

are missing-at-random.[41] Table 7 reports summary statistics borrowed from Hashmi (2013). In particular, one can see that some industries are never granted patents.[42] Table 8 describes the industries present in the sample.

Evidence of Time-Varying Unobserved Heterogeneity. Before estimating a NGFE model, I investigate the existence of a latent clustering structure by applying the pairwise differencing estimator developed in Mugnier (2022) to ABBGH's residuals:

$$\text{cwpatent}_{it} - \widehat{\mathbb{E}}[\text{cwpatent}_{it} \mid \text{comp}_{it}, \widehat{\nu}_i, \widehat{\xi}_t] = \text{cwpatent}_{it} - \exp(\widehat{g}(\text{cwpatent}_{it}) + \widehat{\nu}_i + \widehat{\xi}_t),$$

plotted in Figure 2. This smooth exploration method allows for an unconstrained number of clusters, run in polynomial time, provides a regularization path for the number of groups and estimate time-varying effects without relying on k-means or computing the NGFE which is subject to local minima.[43] Figure 3 and Figure 4 plot the regularization path corresponding to the largest plateau, i.e., for a choice of the regularization parameter such that $\widehat{G} = 3$, and time effects respectively. Figures 4 reveals one cluster with residuals centered around zero and low variance (in red), one cluster with higher volatility and statistically different from zero at several periods and whose CI does not intersect that of the first cluster at least at one period (in blue), and a very high volatility cluster (in green) that consists of industries with missing values. There is evidence of time-varying unobserved heterogeneity.

A Mildly Inverted-U Relationship. I now estimate the following NGFE model:

$$\begin{aligned} &\Pr(\text{cwpatent}_{it} = p \mid \text{comp}_{it}, g_i, \alpha_{g_i t}) \\ &\quad = \frac{\exp(p(g(\text{comp}_{it}) + \alpha_{g_i t}) \exp(-\exp(g(\text{comp}_{it}) + \alpha_{g_i t})))}{p!}, \quad \forall p \in \{0, 1, ...\}, \end{aligned} \tag{25}$$

where $g_i \in \{1, ..., G\}$ is industry i's unknown cluster membership and $(\alpha_{1t}, ..., \alpha_{Gt})' \in \mathbb{R}^G$ are time-specific unobserved effects accounting for unobserved confounding variations in the propensity to patent and product market competition in each of the G clusters. Given the small number of industries, I report results for $G \in \{2, 3, 4\}$. Models (24) and (25) are non-nested as $G << N$.

Table 9 and Figure 5 replicate ABBGH's Table I and Figure I, and additionally show results of NGFE estimation for the choices $G \in \{2, 3, 4\}$, and using $2,000$ random initializers around 0^{2+GT}. Two results are striking. When $G = 2$, the in-sample relationship (no extrapolation) is a significant but mildly increasing relationship. This can be explained by the structure of the cluster effects

[41] While the time dimension is large, the cross-sectional dimension is slightly at odd with the asymptotic framework considered in the paper. Still the economic point applies and it is likely that larger datasets with more digits will be available in the near future.

[42] This does not mean that such industries do not innovate. Patenting is an imperfect measure of innovation in several aspects (Boldrin and Levine, 2013). Many studies perform robustnes checks by using R&D expenses as an alternative measure (Aghion, Bloom, Blundell, Griffith, and Howitt, 2005).

[43] Yet, its statistical guarantees are currently not known in the Poisson model.

discussed in the next paragraph: when $G = 2$, the two estimated clusters do not exhibit a lot of variation over time. Estimation then acts as a constrained classical fixed effect estimator (where industry-specific effects only have two points of support). When G increases, I find strong evidence of a mildly inverted-U relationship. Estimates of the competition parameters are still significantly different from zero but the inverted-U relationship is dramatically less pronounced (the curve is flatter) when unobserved heterogeneity is allowed to be time-varying.

Clustered Unobserved Innovation Dynamics. The 70-90's are characterized by the extremely rapid development of electronics, networks and the Internet. It is likely that economies of scale, shocks and unobserved innovation trends are not the same for each industry. Figure 6 confirms this intuition by plotting the estimated cluster-specific effects obtained in specifications (3)-(5) from Table 9, where the data-driven clustering of industries is displayed in Figure 7.

The NGFE estimates of the unobserved determinants of innovation reveal heterogeneous, time-varying patterns, in particular for $G \geq 3$. Setting $G = 2$ delivers two clusters that experience stable innovation paths over time, albeit at very different levels. Cluster 1, which I refer to as the "high-innovation" cluster, mostly contains highly-patenting, highly-competitive industries. It includes Manufacture of office machinery and data processing equipment, Electrical and electronic engineering, Manufacture of motor vehicles and parts therof, and Manufacture of other transport equipment, but also Chemical industry. Cluster 2, which I refer to as "low-innovation" mostly includes low-patenting, low-competition: metal manufacturing, textile industry, and processing of rubber and plastics, among others. This clustering structure of unobserved heterogeneity is broadly consistent with an additive fixed-effects representation, as the cluster effects $\widehat{\alpha}_{1t}$ and $\widehat{\alpha}_{2t}$ are approximately parallel over time. In contrast, when allowing for more than two clusters, newly estimated clusters are not consistent with a fixed effects model. For $G = 3$, Cluster 2 does not change significantly but the vast majority of industries from Cluster 1 now belongs to Cluster 3 ("steady-catchers") as they experience a steadily increase during the all period towards the unobserved innovation level of Cluster 1. Only the car, food and tobacco, and chemical industries remain in the stable "high-innovation" Cluster 1 whereas Cluster 3 now includes electrical and electronical engineering, office machinery and data processing equipment. Finally, when $G = 4$, Cluster 3 further splits into two neck-to-neck catching-up clusters of industries. The new Cluster 4 ("Noisy-catchers"), which is more volatile in the race, contains other manufacturing industries and transport equipment. Steadily increasing industries now include, among others: Manufacture of office machinery and data processing equipment, and Electrical and electronic engineering.

Figure 8 plots estimated cluster effects, competition and innovation by estimated cluster memberships. It suggests that the relationship between observables and unobservables is complex and hardly predictable from observables only.

Endogeneity. Because competition is likely to be an endogeneous variable, ABBGH use a control function approach by including the residual of a first-stage where the lerner index is predicted by a set of policy instruments such as the Thatcher era privatizations, the EU Single Market Programme, and the Monopoly and Merger Commission investigations at the industry level (see Table II in ABBGH), as an additional regressor in their main specification. The first and fourth columns of Table 10 show that coefficient estimates are similar to Table 9 in the case of NGFE models.

Testing for Structural Break. Finally, I revisit Correa (2012) who tests for the existence of a structural break in 1981. The author finds a decreasing relationship before but no effects of competition afterwards, which would spuriously explain ABBGH's inverted-U relationship. In contrast, a NGFE specification with four clusters shows evidence of a mildly relationship before 1981, but still no significant relationship afterwards (see Table 10).

8 Conclusion

In this paper, I study the nonparametric identification and estimation of a new class of nonlinear panel data models that accomodates clustered patterns of time-varying unobserved heterogeneity. Sufficient low-level conditions delivering identification of all parameters are provided. Because nonparametric estimation might be overwhelmingly cumbersome in panel data with moderate length, I propose semiparametric NGFE estimators that are free of the incidental parameters problem when $T = o(N)$, which sharply contrasts with many competing approaches. Individuals are uniformly classified in the limit as T grows at least as some power of N, and cluster-specific and slope coefficient estimates are asymptotically normal (and centered at the true value). A simple Lloyd's algorithm is shown to perform well in Monte-Carlo simulation. By applying this new estimator to revisit Aghion, Bloom, Blundell, Griffith, and Howitt (2005), I demonstrate that their so-called inverted-U relationship between innovation and product market competition is sensitive to the researcher's choice of whether controling for time-varying grouped effects or not. I document a data-driven clustering of industries. In particular, once controlling for two groups, the relationship becomes increasing. Once controlling for $3 \leq G \leq 4$ clusters, the relationship becomes a mildly inverted-U.

Interesting research avenues include improving computational execution time and developing an estimation approach that would estimate the number of groups with theoretical guarantees (e.g., consistency). In Mugnier (2022), I propose such an estimator for linear versions of NGFE models and obtain this result under relatively weak conditions (see Proposition 3.1). Given such a promising result, it would be nice to extend the approach and prove similar large sample properties for more general nonlinear models, including those considered in this paper. I leave such extensions for future work.

Appendix

A Proof of the Results

I introduce some notation. For any $(a,b) \in \mathbb{R}^2$, I let $a \vee b := \max\{a,b\}$ and $a \wedge b := \min\{a,b\}$. λ denotes the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ collects the Borel sets on $\mathbb{R}$. The abbreviation "a.e." stands for "almost everywhere" (with respect to an appropriate measure). Let $\xrightarrow{d}$ and $\xrightarrow{p}$ denote convergence in distribution and convergence in probability respectively. For any sequence of random variables $\{U_n : n \in \mathbb{N}\}$ such that $U_n \xrightarrow{p} U$, let $\text{plim}_{n\to\infty} U_n := U$. $U_n = O_p(1)$ (resp. $o_p(1)$) means U_n is bounded in probability (resp. converges in probability to zero). $U_n = O_p(R_n)$ means that $U_n = R_n \times V_n$ with $V_n = O_p(1)$; $U_n = o_p(R_n)$ means that $U_n = R_n \times V_n$ with $V_n = o_p(1)$.

A.1 Proof of Theorem 1

Part 1.

Identification of $W_N^0 \in \{0,1\}^{N\times N}$ for all $N \in \mathbb{N}^*$. Let $N \in \mathbb{N}^*$. By Assumption 2, there exist $\mathcal{X}^0 \subset \mathcal{X}$, $\overline{y} \in \mathcal{Y}$, and a known functional ϕ such that, for all $(i,j) \in \mathcal{N}^2$, the (i,j)-th entry of W_N^0, W_{ijN}^0, satisfies $W_{ijN}^0 := \mathbf{1}\left\{g_i^0 = g_j^0\right\} = \phi(\rho_i, \rho_j)$ with $\rho_i(x) : \mathcal{X}^0 \ni x \mapsto \Pr(Y_{i2} = \overline{y} \mid X_{i2} = x, g_i^0, \mu_i^0, \xi_i^0)$. It is then sufficient to show that, for all $i \in \mathcal{N}$, ρ_i is identified. Let $(i,x) \in \mathcal{N} \times \mathcal{X}^0$. Under Assumptions 1(b) and 3(a), and conditional on the σ-algebra generated by $(g_i^0, \mu_{g_i^0}^{0\prime}, \xi_i^{0\prime})'$, the time-series process $\left\{(Y_{it}, X_{it}')' : t \geq 2\right\}$ is strictly stationary strong mixing and satisfies regularity conditions given in Hansen (2008) to obtain consistency of the Nadaraya-Watson estimator of $\mathbb{E}\left[\mathbf{1}\{Y_{it} = \overline{y}\} \mid X_{i2} = x, g_i^0, \mu_{g_i^0}^0, \xi_i^0\right]$. Hence, point identification of $\mathbb{E}\left[\mathbf{1}\{Y_{i2} = \overline{y}\} \mid X_{i2} = x, g_i^0, \mu_{g_i^0}^0, \xi_i^0\right] = \rho_i(x)$ follows by pooling unit i's choices when $(Y_{it}, X_{it}')' \in \{\overline{y}\} \times \mathcal{B}_T(x)$, where $\mathcal{B}_T(x)$ is a well-chosen shrinking neighborhood of x as $T \to \infty$ (e.g., using any well-chosen kernel $\mathcal{K}$ and bandwidth h_T).

Identification of G^0. For any fixed $N \in \mathbb{N}^*$, let R_N^0 denote the number of distinct rows in W_N^0. By the previous paragraph, R_N^0 is identified. But R_N^0, which is also the rank of W_N^0, is exactly the number of clusters represented in the finite sample of size N. Under Assumptions 1(a) and 2(b), $G^0 = \limsup_{N\to\infty} R_N^0$ is thus identified.[44]

Part 2.

Identification of β^0. Let $(i,t) \in \mathbb{N}^{*2}$. By Part 1, $\mathcal{C}^0(i) := \left\{j \in \{1,...,N\} : g_j^0 = g_i^0\right\}$ is identified for all $N \in \mathbb{N}^*$. Under Assumption 1(a) and 2(b), conditional on $(\gamma^{0\prime}, \alpha^{0\prime}, \lambda^{0\prime}, \mu^{0\prime})'$, $\left\{(Y_{jt}, X_{jt}')' : j \in \mathcal{C}^0(i) \setminus \{i\}\right\}$ is an identified infinite sequence of i.i.d. random variables. By applying Theorem 4.1 in Ichimura

[44] From an estimation perpective, one would need conditions on the joint rate of convergence of (N,T) to ensure adequate controls of the error terms (ρ_i should typically be estimated in sup-norm on $\mathcal{X}^0$ at some polynomial rate in T).

(1993) with $\varphi(\cdot) = \sum_{y\in\mathcal{Y}} y h^0(y, \cdot + \alpha^0_{g^0_i t})$, whose conditions 4.1 and 4.2(1-3) hold under Assumptions 1(c) and 3, β^0 is identified up-to-scale. Because $\left\|\beta^0\right\| = 1$, β^0 is identified.

Identification of cluster-specific time effects α^0_{gt} for all $(g,t) \in \mathcal{G}^0 \times \mathbb{N}^*$, up to cluster relabeling. Given identification of W^0_N for all $N \in \mathbb{N}^*$, I build the G^0 groups sequentially starting from $N = 2$, $N = 3$,... and regrouping at each step units with same rows in W^0_N. Without loss of generality, I assume that the resulting labeling matches the true labeling. Let $t \in \mathbb{N}^*$, $x \in \mathcal{X}$, and $\underline{y} \in \mathcal{Y}$ verifying Assumptions 4. By pooling choices of individuals in cluster g and $\widetilde{g}$ at time t for which $Y_{it} = \underline{y}$ and $X_{it} = x$, and applying a standard LLN using Assumptions 1(a) and 1(c), the following probabilities are identified:

$$\Pr\left(Y_{1t} = \underline{y} \mid X_{1t} = x, g^0_1 = g, \alpha^0_{gt}\right) = h^0\left(\underline{y}, x'\beta^0 + \alpha^0_{gt}\right),$$
$$\Pr\left(Y_{1t} = \underline{y} \mid X_{1t} = x, g^0_1 = \widetilde{g}, \alpha^0_{\widetilde{g}t}\right) = h^0\left(\underline{y}, x'\beta^0 + \alpha^0_{\widetilde{g}t}\right).$$

By Assumption 5 (eq. (8)), I can find $x_1, x_2 \in \mathcal{X}$ such that

$$\Pr\left(Y_{1t} = \underline{y} \mid X_{1t} = x_2, g^0_1 = g, \alpha^0_{gt}\right) = \Pr\left(Y_{1t} = \underline{y} \mid X_{1t} = x_1, g^0_1 = \widetilde{g}, \alpha^0_{\widetilde{g}t}\right),$$

or, equivalently,

$$h^0\left(\underline{y}, x_2'\beta^0 + \alpha^0_{gt}\right) = h^0\left(\underline{y}, x_1'\beta^0 + \alpha^0_{\widetilde{g}t}\right). \tag{26}$$

By strict monotonicity of $h^0(\underline{y}, \cdot)$, I can invert (26) and identify $\alpha^0_{\widetilde{g}t} - \alpha^0_{gt} = (x_2 - x_1)'\beta^0$. As β^0 is already identified, it follows that $\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}$ is identified. Because the result holds for all $(g, \widetilde{g}, t)$, it holds for $g = t = 1$ (for which $\alpha^0_{gt} = 0$ by the normalization assumption), thus $(\alpha^0_{g1})_{g\in\mathcal{G}^0}$ is identified. A similar reasoning but now identifying $x_1, x_2 \in \mathcal{X}$ such that eq. (9) holds in place of eq. (8) yields identification of $\alpha^0_{g\widetilde{t}} - \alpha^0_{gt}$ for all $(g, t, \widetilde{t})$, and, in turn, that of $(\alpha^0_{1t})_{t\in\mathbb{N}^*}$. Identification of α^0_{gt} for all (g,t) then follows because, for all (g,t) with $g \neq 1$ and $t \neq 1$, α^0_{gt} can be decomposed as

$$\alpha^0_{gt} = \underbrace{\alpha^0_{gt} - \alpha^0_{1t}}_{:=a} + \underbrace{\alpha^0_{1t}}_{:=b},$$

where a and b are identified. Finally, $h^0(y, z)$ is identified as a function of $y \in \mathcal{Y}$ and index $z = X_{it}'\beta^0 + \alpha^0_{g^0_i t}$.
The proof of Theorem 1 is complete.

A.2 Sufficient Condition for Assumption 2(a)

Consider the following assumption.

Assumption 10

(a) There exists an open set $\mathcal{X}^1 \subset \mathcal{X}$ such that, for all $(i, j, g, \tilde{g}, x) \in \mathbb{N}^{*2} \times \mathcal{G}^{02} \times \mathcal{X}^1$, the conditional distribution $\alpha^0_{g2} \mid X_{i2} = x, g^0_i = g, \mu^0_{g^0_i}, \xi^0_i$ admits a fully supported density $f_{\alpha^0_{g2}|X_{i2}=x, g^0_i=g, \mu^0_g, \xi^0_i}(\alpha)$

with respect to the Lebesgue measure such that

$$f_{\alpha^0_{g2}|X_{i2}=x,g^0_i=g,\mu^0_g,\xi^0_i}(\alpha) = f_{\alpha^0_{\widetilde{g}2}|X_{j2}=x,g^0_j=\widetilde{g},\mu^0_{\widetilde{g}},\xi^0_j}(\alpha), \quad \lambda(\alpha)\text{-}a.e.$$

if and only if $g = \widetilde{g}$.

(b) There exists $k \in \{1, ..., p\}$ *such that* $\beta^0_k \neq 0$ *and* $X_{i2,k} \perp\!\!\!\perp \alpha^0_{g^0_i 2} \mid X_{i2,(-k)}, g^0_i, \mu^0_{g^0_i}, \xi^0_i$. *Moreover, almost surely,* $\text{Supp}\left(X_{i2,k} \mid X_{i2,(-k)}, g^0_i, \mu^0_{g^0_i}, \xi^0_i\right)$ *is open.*

(c) There exists $y \in \mathcal{Y}$ *such that* $\psi_y : v \mapsto h^0(y, v)$ *is stricly monotonic, real analytic with bounded first derivative* ψ'_y *such that* $\int \left|\psi'_y\right| d\lambda < \infty$.[45] *Moreover, the characteristic function of* ζ *with density* $f_\zeta(z) = \frac{\left|\psi'_y(z)\right|}{\int \left|\psi'_y\right| d\lambda}$ *does not vanish and is infinitely often differentiable in* $\mathbb{R}\backslash A$ *for some set* A *such that* $\lambda(A) = 0$.

Assumption 10(b) requires the existence of a special regressor (as in Honoré and Lewbel, 2002), but (possibly) without large support (it depends on the support of the unobserved heterogeneity). Assumption 10(c) imposes smoothness conditions including real-analyticity of the link functions. Example of distributions satisfying these are given in, e.g., D'Haultfoeuille (2010). Real-analyticity can be relaxed to continuous differentiability by strenghtenning the support in Assumption 10(b) to be the full real line, which is equivalent to having a special regressor with large support à la Honoré and Lewbel (2002).

Lemma 1 *If Assumptions 1(c) and 10 hold, then Assumption 2(a) holds.*

Proof of Lemma 1 W.l.o.g. I assume that $k = 1$ and denote $x_{(-1)} = (x_j)_{j\in\{2,...,p\}}$. Let $x = (x_1, x'_{(-1)})' \in \mathcal{X}^1$, and $y \in \mathcal{Y}$ verifying Assumption 10(c). I proceed in two steps. In the first step, I construct $\mathcal{X}^0 \subset \mathcal{X}^1$. In the second step, I construct ϕ that fulfills Assumption 2.
Step 1: Let $(i, x) \in \mathcal{N} \times \mathcal{X}^1$ and $\rho_i(x) := \Pr\left(Y_{i2} = y \mid X_{i2} = x, g^0_i, \mu^0_{g^0_i}, \xi^0_i\right)$. By the law of total expectations, Assumption 1(c), using equation (2), and Assumption 10(a), I obtain

$$\begin{aligned}
\rho_i(x) &= \mathbb{E}\left[\Pr\left(Y_{i2} = y \mid X_{i2} = x, g^0_i, \alpha^0, \lambda^0, \mu^0, \xi^0\right) \mid X_{i2} = x, g^0_i, \mu^0_{g^0_i}, \xi^0_i\right] \\
&= \mathbb{E}\left[\Pr\left(Y_{i2} = y \mid X_{i2} = x, g^0_i, \alpha^0_{g^0_i 2}\right) \mid X_{i2} = x, g^0_i, \mu^0_{g^0_i}, \xi^0_i\right] \\
&= \mathbb{E}\left[\psi_y\left(x'\beta^0 + \alpha^0_{g^0_i 2}\right) \mid X_{i2} = x, g^0_i, \mu^0_{g^0_i}, \xi^0_i\right] \\
&= \int_{\mathbb{R}} \psi_y\left(x'\beta^0 + \alpha\right) f_{\alpha^0_{g^0_i 2}|X_{i2}=x,g^0_i,\mu^0_{g^0_i},\xi^0_i}(\alpha) d\lambda(\alpha).
\end{aligned} \tag{27}$$

[45] Let $I \subset \mathbb{R}$ be an open set. A function $f : I \to \mathbb{R}$ is called "analytic" if for any $x_0 \in I$ there is a neighborhood J of x_0 and a power series $\sum a_n(x - x_0)^n$ such that $f(x) = \sum_n a_n(x - x_0)^n \quad \forall x \in J$ (see, e.g., Krantz and Parks, 2002).

By Assumption 10(b), there exists $\epsilon > 0$ and an open set $\mathcal{X}^0 = \{x + (v, 0')' : v \in (-\epsilon, \epsilon)\} \subset \mathcal{X}^1$ with $\Pr(X_{i2} \in \mathcal{X}^0) > 0$ such that, for all $w \in \mathcal{X}^0$, almost everywhere $f_{\alpha^0_{g^0_i 2}|X_{i2}=w, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha) = f_{\alpha^0_{g^0_i 2}|X_{i2}=x, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha)$. Since $\mathcal{X}^0 \subset \mathcal{X}^1$, eq. (27) yields, for all $w \in \mathcal{X}^0$,

$$\rho_i(w) = \int_{\mathbb{R}} \psi_y\left(w'\beta^0 + \alpha\right) f_{\alpha^0_{g^0_i 2}|X_{i2}=x, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha) d\lambda(\alpha).$$

By Assumption 10(c), $w \mapsto \rho_i(w)$ is differentiable on $\mathcal{X}^0$ and, for all $w \in \mathcal{X}^0$,

$$\begin{aligned} \left.\frac{\partial \rho_i(z_1, ..., z_p)}{\partial z_1}\right|_{z=w} &= \beta^0_1 \int_{\mathbb{R}} \psi'_y\left(w'\beta^0 + \alpha\right) f_{\alpha^0_{g^0_i 2}|X_{i2}=x, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha) d\lambda(\alpha) \\ &= \beta^0_1 \left(1 - 2\mathbf{1}\left\{\psi'_y(0) < 0\right\}\right) \int_{\mathbb{R}} \left|\psi'_y\left(w'\beta^0 + \alpha\right)\right| f_{\alpha^0_{g^0_i 2}|X_{i2}=x, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha) d\lambda(\alpha), \end{aligned} \tag{28}$$

where the second equality follows from the strict monotonicity of $\psi_y(\cdot)$.

Step 2: Let $\Delta(a, b) := a - b$ and ∂_1 be the partial differencing operator with respect to the first argument (for multivalued functions). I prove below that $\phi(f, g) := \mathbf{1}\{\Delta(\partial_1 f, \partial_1 g) = 0\}$ verifies Assumption 2(a). I have to show that, for all $(i, j) \in \mathcal{N}^2$,

$$\left.\frac{\partial \rho_i(z_1, ..., z_p)}{\partial z_1}\right|_{z=w} = \left.\frac{\partial \rho_j(z_1, ..., z_p)}{\partial z_1}\right|_{z=w} \quad \forall w \in \mathcal{X}^0 \iff g^0_i = g^0_j. \tag{29}$$

Let $(i, j) \in \mathcal{N}^2$.

$\underline{\Longleftarrow:}$ Suppose that $g^0_j = g^0_i$ and let $w \in \mathcal{X}^0$. By Assumption 10(c), I have

$$f_{\alpha^0_{g^0_i 2}|X_{i2}=x, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha) = f_{\alpha^0_{g^0_j 2}|X_{j2}=x, g^0_j, \mu^0_{g^0_j}, \xi^0_j}(\alpha), \quad \lambda(\alpha) - \text{a.e.}.$$

Equation (28) then implies $\left.\frac{\partial \rho_i(z_1, ..., z_p)}{\partial z_1}\right|_{z=w} = \left.\frac{\partial \rho_j(z_1, ..., z_p)}{\partial z_1}\right|_{z=w}$.

$\underline{\Longrightarrow:}$ Suppose that, for all $w \in \mathcal{X}^0$,

$$\left.\frac{\partial \rho_i(z_1, ..., z_p)}{\partial z_1}\right|_{z=w} = \left.\frac{\partial \rho_j(z_1, ..., z_p)}{\partial z_1}\right|_{z=w}.$$

Dividing each side of this equation by $\int \left|\psi'_y\right| d\lambda > 0$, using (28) and the fact that

$$\left|\left(1 - 2\mathbf{1}\left\{\psi'_y(0) < 0\right\}\right) \beta^0_1\right| = \left|\beta^0_1\right| > 0,$$

I obtain, denoting $f_{\alpha^0_{g^0_i}}(\alpha) := f_{\alpha^0_{g^0_i 2}|X_{i2}, g^0_i, \mu^0_{g^0_i}, \xi^0_i}(\alpha)$, for all $w \in \mathcal{X}^0$,

$$\int_{\mathbb{R}} f_\zeta\left(w'\beta^0 + \alpha\right) f_{\alpha^0_{g^0_i}}(\alpha) d\lambda(\alpha) = \int_{\mathbb{R}} f_\zeta\left(w'\beta^0 + \alpha\right) f_{\alpha^0_{g^0_j}}(\alpha) d\lambda(\alpha).$$

I show below that this constraint is equivalent to $f_{\alpha^0_{g^0_j}} = f_{\alpha^0_{g^0_i}}$ a.e., which, by Assumption 10(a), in turn implies $g^0_i = g^0_j$. Specifically, I show that the solution set $\mathcal{S}^* \subset L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ to the integral inverse problem: $f_\alpha \in \mathcal{S}^*$ if and only if

$$\int_{\mathbb{R}} f_\zeta \left(w'\beta^0 + \alpha\right) f_{\alpha^0_{g^0_i}}(\alpha) d\lambda(\alpha) = \int_{\mathbb{R}} f_\zeta \left(w'\beta^0 + \alpha\right) f_\alpha(\alpha) d\lambda(\alpha) \quad \forall w \in \mathcal{X}^0, \tag{30}$$

verifies $\mathcal{S}^* = \left\{ f \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda) : f_\alpha = f_{\alpha^0_{g^0_i}} \text{ a.e.} \right\}$. Suppose $f^*_\alpha \in \mathcal{S}^*$ and consider the change of variable $z = w'\beta^0 + \alpha$ in (30). Then, for all $\delta \in (x'\beta^0 - \beta^0_1\epsilon, x'\beta^0 + \beta^0_1\epsilon) \subset \mathbb{R}$,

$$\int_{\mathbb{R}} f_\zeta(z) f_{-\alpha^0_{g^0_i}}(\delta - z) d\lambda(z) = \int_{\mathbb{R}} f_\zeta(z) f^*_{-\alpha}(\delta - z) d\lambda(z). \tag{31}$$

Note that both sides of eq. (31) are convolutions of f_ζ with $df_{-\alpha^0_{g^0_i}}$ or $df^*_{-\alpha}$. By letting

$$\mathcal{W} : \delta \mapsto \int_{\mathbb{R}} f_\zeta(\delta - z) \left[f_{-\alpha^0_{g^0_i}}(z) - f^*_{-\alpha}(z) \right] d\lambda(z),$$

and using commutativity of the convolution product, eq. (31) implies that there exists an open set $U \subset \mathbb{R}$ such that

$$\mathcal{W}(\delta) = 0, \quad \forall \delta \in U. \tag{32}$$

Given Assumption 10(c), it can be shown that $\mathcal{W} : \mathbb{R} \to \mathbb{R}$ is real-analytic (see footnote 45). A continuation theorem for real analytic functions (see e.g. Corollary 1.2.5 in Krantz and Parks, 2002) implies that eq. (32) holds for all $\delta \in \mathbb{R}$, i.e.:

$$\int_{\mathbb{R}} f_\zeta(\delta - z) \left[f_{-\alpha^0_{g^0_i}}(z) - f^*_{-\alpha}(z) \right] d\lambda(z) = 0, \quad \forall \delta \in \mathbb{R}. \tag{33}$$

Since the functions $f_\zeta, f_{-\alpha^0_{g^0_i}}$, and $f^*_{-\alpha}$ belong to $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, I can apply Fourrier transformation on both sides of eq. (33) to obtain

$$\varphi_{f_\zeta}(v) \times \left[\varphi_{f_{-\alpha^0_{g^0_i}}}(v) - \varphi_{f^*_{-\alpha}}(v) \right] = 0, \quad \forall v \in \mathbb{R}, \tag{34}$$

where φ_f is the Fourrier transform of f. By Assumption 10(c) again, the set

$$\{v \in \mathbb{R} : \varphi_\zeta(v) = 0\}$$

is of zero Lebesgue measure. Equation (34) therefore implies $\varphi_{f_{-\alpha^0_{g^0_i}}} = \psi_{f^*_{-\alpha}}$ a.e.. Since Fourrier transforms are continuous, I obtain $\varphi_{f_{-\alpha^0_{g^0_i}}} = \varphi_{f^*_{-\alpha}}$ everywhere and thus $f_{\alpha^0_{g^0_i}} = f^*_\alpha$ everywhere.

The proof of Lemma 1 is complete.

A.3 Proof of Theorem 2

The key argument is to linearize problem (18) by mean of a second-order Taylor expansion, bounding the log-likelihood function by below by a quadratic function similar to that appearing in Lemma A.2 in Bonhomme and Manresa (2015). For all $\theta = (\beta', \alpha', \gamma')' \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T}$, define

$$\widehat{Q}(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} - \ln\left(\Psi\left(Q_{it} Z_{it}\right)\right),$$

where $Z_{it} = X_{it}'\beta + \alpha_{g_i t}$ and $Q_{it} = 2Y_{it} - 1$. Note that Z_{it} is an implicit function of θ but I drop this conditioning for the sake of clarity and let $Z_{it}^0 = X_{it}'\beta^0 + \alpha^0_{g_i^0 t}$ denote Z_{it} evaluated at the true parameter value θ^0. Note that the NGFE estimator $\widehat{\theta}$ minimizes $\widehat{Q}(\cdot)$ over all $\theta \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T}$. Define the auxiliary quadratic function:

$$\check{Q}(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(X_{it}'\left(\beta - \beta^0\right) + \alpha_{g_i t} - \alpha^0_{g_i^0 t}\right)^2,$$

and let $\overline{z} := \sup_{(\beta', \alpha', g, x)' \in \mathcal{B} \times \mathcal{A}^{\mathcal{G}^0 T} \times \mathcal{G}^0 \times \cup_{t=1,\ldots,i=1,\ldots} \mathrm{Supp}(X_{it})} |Z_{it}|$ and $\mathcal{Z} = [-\overline{z}, \overline{z}]$. Note that $\mathcal{Z}$ is a well-defined segment of $\mathbb{R}$ by Assumptions 7(a) and 7(b). By second-order Taylor expansion, for any z_1, z_2 in $\mathcal{Z}$,

$$-\ln \Psi(z_1) = -\ln \Psi(z_2) - (\ln \Psi)'(z_2)(z_1 - z_2) - \frac{1}{2}(\ln \Psi)''(z^*)(z_1 - z_2)^2,$$

for some $z^* \in]z_1 \wedge z_2, z_1 \vee z_2[$. By continuity of $z \mapsto -(\ln \Psi)''(z)$ and because $-(\ln \Psi)''(z) > 0$ by Assumption 6(b), there exists a constant $b_{\min} > 0$ such that, for all $z \in \mathcal{Z}$,

$$b_{\min} \leq -(\ln \Psi)''(z).$$

Hence, for all $z_1, z_2 \in \mathcal{Z}$

$$-\ln \Psi(z_1) \geq -\ln \Psi(z_2) + s(z_2)(z_1 - z_2) + \frac{b_{\min}}{2}(z_1 - z_2)^2, \tag{35}$$

where $s(z) = -(\ln \Psi)'(z)$. Now substitute $Q_{it}Z_{it}$ for z_1 and $Q_{it}Z_{it}^0$ for z_2, and averaging (35) over i, t, I have, for all $\theta \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0 T}$,

$$\widehat{Q}(\theta) - \widehat{Q}(\theta^0) \geq \frac{b_{\min}}{2} \check{Q}(\theta) + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E_{it}\left(Q_{it}\left(Z_{it} - Z_{it}^0\right)\right), \tag{36}$$

where $E_{it} = s\left(Q_{it}Z_{it}^0\right)$. Since the estimated parameter $\widehat{\theta}$ minimizes $\widehat{Q}(\cdot)$, deduce

$$0 \geq \widehat{Q}(\widehat{\theta}) - \widehat{Q}(\phi^0) \geq \frac{b_{\min}}{2}\check{Q}(\widehat{\theta}) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E_{it}\left(Q_{it}\left(\widehat{Z}_{it} - Z_{it}^0\right)\right), \tag{37}$$

where $\widehat{Z}_{it} = X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}$. I start by showing the following uniform convergence result, which is reminiscent of Lemma A.1 in Bonhomme and Manresa (2015).

Lemma 2 *Let Assumption 6 and Assumptions 7(a)-(b) hold. Then,*

$$\sup_{\theta \in \mathcal{B}\times\mathcal{G}^{0N}\times\mathcal{A}^{G^0 T}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E_{it}\left(Q_{it}\left(Z_{it} - Z_{it}^0\right)\right) = o_p(1).$$

Proof of Lemma 2: The proof closely follows that of Lemma A.1 in Bonhomme and Manresa (2015), up to a few adjustments.

$$\begin{aligned}
&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E_{it}\left(Q_{it}\left(Z_{it} - Z_{it}^0\right)\right)\\
&= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Q_{it}E_{it}\left(X_{it}'\left(\beta - \beta^0\right) + \alpha_{g_i t} - \alpha^0_{g_i^0 t}\right)\\
&= \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Q_{it}E_{it}X_{it}\right)'\left(\beta - \beta^0\right) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E_{it}Q_{it}\alpha_{g_i t} - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} E_{it}Q_{it}\alpha^0_{g_i^0 t}.
\end{aligned}$$

Let $\mathcal{F}_t = \sigma\left(\left\{\gamma^0, \alpha^0, \mathbf{X}_-^{(t)}, \boldsymbol{\varepsilon}_-^{(t-1)}\right\}\right)$ denote the σ-field generated by $\gamma^0, \alpha^0, \mathbf{X}_-^{(t)}$, and $\boldsymbol{\varepsilon}_-^{(t-1)}$. Under Assumptions 6(a) and 6(b), for all $s < t$, I have

$$\begin{aligned}
\mathbb{E}\left(Q_{it}Q_{is}E_{it}E_{is}X_{it}'X_{is}\right) &= \mathbb{E}\left(\mathbb{E}(Q_{it}Q_{is}E_{it}E_{is}X_{it}'X_{is} \mid \mathcal{F}_t)\right)\\
&= \mathbb{E}\left(X_{it}'X_{is}Q_{is}E_{is}\mathbb{E}(Q_{it}E_{it} \mid \mathcal{F}_t)\right)\\
&= \mathbb{E}\left(X_{it}'X_{is}Q_{is}E_{is}\mathbb{E}\left(\frac{Y_{it} - \Psi(Z_{it}^0)}{\Psi(Z_{it}^0)(1 - \Psi(Z_{it}^0))}\Psi'(Z_{it}^0) \mid \mathcal{F}_t\right)\right)\\
&= \mathbb{E}\left(X_{it}'X_{is}Q_{is}E_{is}\underbrace{\frac{\mathbb{E}(Y_{it} - \Psi(Z_{it}^0) \mid \mathcal{F}_t)}{\Psi(Z_{it}^0)(1 - \Psi(Z_{it}^0))}\Psi'(Z_{it}^0)}_{=0}\right)\\
&= 0,
\end{aligned}$$

where the penultimate equality follows because $\Psi'(Z_{it}^0)$ is $\mathcal{F}_t$-measurable, and the last equality follows from $\mathbb{E}(Y_{it} \mid \mathcal{F}_t) = \Psi(Z_{it}^0)$. By Cauchy-Schwarz (CS) inequality, and using Assumption 6(b), 7(b) and $Q_{it}^2 = 1$, there exists a constant $M' > 0$ such that, for $s = t$,

$$\mathbb{E}\left(Q_{it}Q_{is}E_{it}E_{is}X_{it}'X_{is}\right) = \mathbb{E}\left(E_{it}^2\left\|X_{it}\right\|^2\right) \leq \sqrt{\mathbb{E}\left(E_{it}^4\right)\mathbb{E}\left(\left\|X_{it}\right\|^4\right)} \leq M' < \infty.$$

Hence, I have

$$\left| \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E} \left(Q_{it} Q_{is} E_{it} E_{is} X_{it}' X_{is} \right) \right| \leq M'. \tag{38}$$

By (38), I have

$$\mathbb{E} \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \frac{1}{T} \sum_{t=1}^{T} Q_{it} E_{it} X_{it} \right\|^2 \right) \leq \frac{M'}{T},$$

so it follows from the Markov inequality that

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Q_{it} E_{it} X_{it} = o_p(1).$$

In addition, $\left\| \beta - \beta^0 \right\|$ is bounded under Assumption 7(a), hence

$$\left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Q_{it} E_{it} X_{it} \right)' (\beta - \beta^0) = o_p(1).$$

I next show that $\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Q_{it} E_{it} \alpha_{g_i t}$ is $o_p(1)$, uniformly on the parameter space. This will imply that $\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Q_{it} E_{it} \alpha^0_{g_i^0 t} = o_p(1)$. I have

$$\begin{aligned} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} Q_{it} E_{it} \alpha_{g_i t} &= \sum_{g \in \mathcal{G}^0} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{1} \left\{ g_i = g \right\} Q_{it} E_{it} \alpha_{gt} \right] \\ &= \sum_{g \in \mathcal{G}^0} \left[\frac{1}{T} \sum_{t=1}^{T} \alpha_{gt} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ g_i = g \right\} Q_{it} E_{it} \right) \right]. \end{aligned}$$

Moreover, by the CS inequality and for all $g \in \mathcal{G}^0$:

$$\left(\frac{1}{T} \sum_{t=1}^{T} \alpha_{gt} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ g_i = g \right\} Q_{it} E_{it} \right) \right)^2 \leq \left(\frac{1}{T} \sum_{t=1}^{T} \alpha_{gt}^2 \right) \times \left(\frac{1}{T} \sum_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ g_i = g \right\} Q_{it} E_{it} \right)^2 \right),$$

where, by Assumption 7(a), $\frac{1}{T} \sum_{t=1}^{T} \alpha_{gt}^2$ is uniformly bounded. Now, note that

$$\begin{aligned} \frac{1}{T} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ g_i = g \right\} Q_{it} E_{it} \right)^2 &= \frac{1}{TN^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{1} \left\{ g_i = g \right\} \mathbf{1} \left\{ g_j = g \right\} \sum_{t=1}^{T} Q_{it} Q_{jt} E_{it} E_{jt} \\ &\leq \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \frac{1}{T} \sum_{t=1}^{T} Q_{it} Q_{jt} E_{it} E_{jt} \right| \\ &\leq \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(Q_{it} Q_{jt} E_{it} E_{jt}) \right| \\ &\quad + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \frac{1}{T} \sum_{t=1}^{T} \left(Q_{it} Q_{jt} E_{it} E_{jt} - \mathbb{E}(Q_{it} Q_{jt} E_{it} E_{jt}) \right) \right|. \end{aligned}$$

Since $\mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt}) = 0$ for $i \neq j$, there exists a constant $M'' > 0$ such that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt})\right| \leq M'' < \infty,$$

and, therefore, $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt})\right| \leq \frac{M''}{N}$. Moreover, by the CS inequality,

$$\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}(Q_{it}Q_{jt}E_{it}E_{jt} - \mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt}))\right|\right)^2$$
$$\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(Q_{it}Q_{jt}E_{it}E_{jt} - \mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt}))\right)^2. \tag{39}$$

Similarly again, I can show that there exists a constant $M''' > 0$ such that

$$\left|\frac{1}{N^2T}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}(Q_{it}Q_{jt}E_{it}E_{js}, Q_{is}q_{js}E_{is}E_{js})\right| \leq M''' < \infty.$$

Hence, the term in the right-hand side of (39) is bounded in expectation by M'''/T. This shows that $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}Q_{it}E_{it}\alpha_{g_it}$ is uniformly $o_p(1)$, and ends the proof of Lemma 2. □

Next, by Lemma A.2 in Bonhomme and Manresa (2015), it follows that

$$\check{Q}(\widehat{\theta}) \geq \widehat{\rho}\left\|\widehat{\beta} - \beta^0\right\|^2, \tag{40}$$

where $\mathrm{plim}_{N,T\to\infty}\widehat{\rho} = \rho > 0$. Hence, combining (37), Lemma 2, and (40) I obtain

$$0 \geq \frac{b_{\min}\rho}{2}\left\|\widehat{\beta} - \beta^0\right\|^2 + o_p(1),$$

from which it is concluded that $\widehat{\beta} = \beta^0 + o_p(1)$.

Lastly, to show convergence in quadratic mean of the estimated unit-specific effects, note that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widehat{g}_it} - \alpha^0_{g^0_it}\right)^2$$
$$= \check{Q}(\theta) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X'_{it}\left(\beta^0 - \widehat{\beta}\right)X'_{it}\left(\beta^0 - \widehat{\beta}\right) - \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}X'_{it}\left(\beta^0 - \widehat{\beta}\right)\left(\alpha^0_{g^0_it} - \widehat{\alpha}_{\widehat{g}_it}\right)$$
$$\leq \check{Q}(\theta) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|X_{it}\right\|^2 \times \left\|\beta^0 - \widehat{\beta}\right\|^2$$
$$+ \left(4\sup_{\alpha\in\mathcal{A}}|\alpha|\right) \times \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|X_{it}\right\| \times \left\|\beta^0 - \widehat{\beta}\right\|,$$

which is $o_p(1)$ by Assumptions 7(a)-7(b), by consistency of $\widehat{\beta}$, and because Lemma 2 and (37)

together imply $\check{Q}(\widehat{\theta}) = o_p(1)$.
This completes the proof of Theorem 2.

A.4 Proof of Theorem 3

A.4.1 Step 1: A Useful Asymptotic Equivalence

Lemma 7 below provides an asymptotic equivalence result which is key to prove Theorem 3. I first prove three lemmas (3, 4, and 5) that help in showing that NGFE estimators achieve uniformly consistent classification of individuals (Lemma 6). This, in turn, allows me to prove Lemma 7.

First, consistency of $\widehat{\alpha}$ for α^0 can be established as in Bonhomme and Manresa (2015). Because the objective function is invariant to relabeling of the cluster labels, the consistency result holds with respect to the Hausdorff distance d_H in $\mathbb{R}^{G^0 T}$, defined by

$$d_H(a,b)^2 = \max\left\{\max_{g\in\mathcal{G}^0}\left(\min_{\widetilde{g}\in\mathcal{G}^0}\frac{1}{T}\sum_{t=1}^{T}\left(a_{\widetilde{g}t}-b_{gt}\right)^2\right), \max_{\widetilde{g}\in\mathcal{G}^0}\left(\min_{g\in\mathcal{G}^0}\frac{1}{T}\sum_{t=1}^{T}\left(a_{\widetilde{g}t}-b_{gt}\right)^2\right)\right\}.$$

Lemma 3 *Let Assumptions 6-7, and 8(a)-8(b) hold. Then, as N and T tend to infinity,*

$$d_H\left(\widehat{\alpha},\alpha^0\right)\xrightarrow{p}0.$$

Proof of Lemma 3: Given Theorem 2, the proof is identical to that of Lemma B.3 in Bonhomme and Manresa (2015). □

Second, I rely on the use of exponential inequalities for dependent processes. Lemma 4 and Lemma 5 are direct consequences of Theorem 6.2 in Rio (2000) (see also Merlevède, Peligrad, and Rio, 2011) and Theorem 3.2 in Lesigne and Volný (2001), respectively.

Lemma 4 (Bonhomme and Manresa (2015), Lemma B.5) *Let z_t be a strongly mixing process with zero mean, with strong mixing coefficient $\alpha[t] \leq \exp(-at^{d_1})$, and tail probabilities $\Pr(|z_t| < z) \leq \exp\left(1-\left(\frac{z}{b}\right)^{d_2}\right)$, where a, b, d_1, and d_2 are positive constants. Then, for all $z > 0$, for all $\delta > 0$,*

$$T^{\delta}\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}z_t\right|\geq z\right)\to 0,\ \ \textit{as } T\to\infty.$$

Lemma 5 [46] *Let $\{z_t,\mathcal{F}_t\}_{t=1}^{T}$ be a martingale difference sequence and assume that there exists $\delta, M > 0$ such that $E(\exp(\delta\,|z_t|)) \leq M$ for all $t = 1,...,T$. Then, for $a > 0$, there exist positive constants*

[46] I found this result in a 2013 unpublished manuscript by A.-B. Kock entitled "Oracle inequalities and variable selection in high-dimensional panel data models" (Lemma 2). For completeness, I report the original proof of the author here.

A and B such that for all $z \geq a/\sqrt{T}$

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \leq A\exp\left(-B(z^2T)^{1/3}\right).$$

Proof of Lemma 5: In the proof of their Theorem 3.2 Lesigne and Volný (2001) show that if $E(\exp(|z_t|) \leq M$ for all $t = 1, ..., T$, then for any $x > 0$ and $t \in (0,1)$, I have

$$\begin{aligned}
&\Pr\left(\left|\sum_{t=1}^{T} z_t\right| > Tz\right) \\
&< \left(2 + \frac{M}{(1-t)^2}\left[\frac{1}{4}t^{4/3}(z^{-2}T^{-1})^{1/3} + t^{2/3}(z^{-2}T^{-1})^{2/3} + 2z^{-2}T^{-1}\right]\right) \\
&\quad \times \exp\left(-\frac{1}{2}t^{2/3}(z^2T)^{1/3}\right). \qquad (41)
\end{aligned}$$

Note that $\Pr\left(\left|\sum_{t=1}^{T} z_t\right| > Tz\right) = \Pr\left(\left|\sum_{t=1}^{T}(\delta z_t)\right| > T(\delta z)\right)$ where $\{\delta z_t\}_{1\leq t\leq T}$, by assumption now satisfy the conditions of Theorem 3.2 in Lesigne and Volný (2001) and so replacing z by δz in (41) yields

$$\begin{aligned}
&\Pr\left(\left|\sum_{t=1}^{T} z_t\right| > Tz\right) \\
&< \left(2 + \frac{M}{(1-t)^2}\left[\frac{1}{4}t^{4/3}\delta^{-2/3}(z^{-2}T^{-1})^{1/3} + t^{2/3}\delta^{-4/3}(z^{-2}T^{-1})^{2/3} + 2\delta^{-2}z^{-2}T^{-1}\right]\right) \\
&\quad \times \exp\left(-\frac{1}{2}t^{2/3}\delta^{2/3}(z^2T)^{1/3}\right).
\end{aligned}$$

Restricting z to be greater than $a/\sqrt{T}$, implying that $z^{-2}T^{-1} \leq 1/a^2$, and using that M, t and δ are constants the conclusion of the lemma follows. □

I am now in position to prove Lemma 6 which extends Lemma B.4 in Bonhomme and Manresa (2015) and shows that $\widehat{g}_i(\beta, \alpha)$ achieves uniformly consistent classification of individuals over a neighbourhood of the true parameter values (β^0, α^0). Note that by the same arguments as in the proof of Lemma B.3 in Bonhomme and Manresa (2015), there exists a permutation $\sigma : \mathcal{G}^0 \to \mathcal{G}^0$ such that

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(g)t} - \alpha_{gt}^0\right)^2 \xrightarrow{p} 0. \qquad (42)$$

By simple relabeling of the elements of $\widehat{\alpha}$, I may take $\sigma(g) = g$. I adopt this convention in the rest of the proof. For any $\eta > 0$, I let $\mathcal{N}_\eta$ denote the set of parameters $(\beta, \alpha) \in \mathcal{B} \times \mathcal{A}^{G^0T}$ that satisfy $\left\|\beta - \beta^0\right\|^2 < \eta$ and $\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt} - \alpha_{gt}^0\right)^2 < \eta$ for all $g \in \mathcal{G}^0$.

Lemma 6 *For $\eta > 0$ small enough, I have, for all $\delta > 0$ and as N and T tend to infinity,*

$$\sup_{(\beta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i(\beta,\alpha) \neq g_i^0\right\} = o_p(T^{-\delta}).$$

Proof of Lemma 6: Note that, from the definition of $\widehat{g}_i(\cdot)$, for all $g \in \mathcal{G}^0$,

$$\mathbf{1}\left\{\widehat{g}_i(\beta,\alpha)=g\right\} \leq \mathbf{1}\left\{\sum_{t=1}^{T} \ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+\alpha_{g_i^0 t}\right)\right)\right) \leq \sum_{t=1}^{T} \ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+\alpha_{gt}\right)\right)\right)\right\},$$

so

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i(\beta,\alpha)\neq g_i^0\right\} &= \sum_{g\in\mathcal{G}^0}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0\neq g\right\}\mathbf{1}\left\{\widehat{g}_i(\beta,\alpha)=g\right\} \\ &\leq \sum_{g\in\mathcal{G}^0}\frac{1}{N}\sum_{i=1}^{N}W_{ig}(\beta,\alpha),\end{aligned}$$

where

$$W_{ig}(\beta,\alpha)=\mathbf{1}\left\{g_i^0\neq g\right\}\times\mathbf{1}\left\{\sum_{t=1}^{T} \ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+\alpha_{g_i^0 t}\right)\right)\right) \leq \sum_{t=1}^{T} \ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+\alpha_{gt}\right)\right)\right)\right\}.$$

I start bounding $W_{ig}(\beta,\alpha)$, for all $(\beta,\alpha)\in\mathcal{N}_\eta$, by a quantity that does not depend on (β,α). To proceed first note that, by Assumption 6(b), and 7(a)-7(b), $v\mapsto\ln\left(\Psi\left(Q_{it}\left(X_{it}'v+\alpha_{gt}\right)\right)\right)$ is uniformly Lipschitz over $(i,t,\alpha,g)\in\{1,...,N\}\times\{1,...,T\}\times A^{G^0T}\times\mathcal{G}^0$, i.e., there exists a constant $L_\beta>0$ such that, for all $(i,t,\alpha,g)\in\{1,...,N\}\times\{1,...,T\}\times A^{G^0T}\times\mathcal{G}^0$, all $\beta_1,\beta_2\in\mathcal{B}$, almost surely

$$\left|\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta_1+\alpha_{gt}\right)\right)\right)-\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta_2+\alpha_{gt}\right)\right)\right)\right|\leq L_\beta\left\|\beta_1-\beta_2\right\|. \tag{43}$$

Similarly, $a\mapsto\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+a\right)\right)\right)$ is uniformly Lipschitz over $(i,t,\beta)\in\{1,...,N\}\times\{1,...,T\}\times\mathcal{B}$, i.e., there exists a constant $L_\alpha>0$ such that, for all $(i,t,\beta)\in\{1,...,N\}\times\{1,...,T\}\times\mathcal{B}$, all $a,b\in\mathcal{A}$, almost surely

$$\left|\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+a\right)\right)\right)-\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+b\right)\right)\right)\right|\leq L_\alpha\left|a-b\right|. \tag{44}$$

Then, by choosing $g=g_i^0,\beta_1=\beta^0$ and $\beta_2=\beta$ in (43), I have, for all (β,α) and all i,

$$\begin{aligned}W_{ig}(\beta,\alpha)\leq{}&\mathbf{1}\left\{g_i^0\neq g\right\}\\ &\times\mathbf{1}\left\{\sum_{t=1}^{T}\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta^0+\alpha_{g_i^0 t}\right)\right)\right)\leq\sum_{t=1}^{T}\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+\alpha_{gt}\right)\right)\right)+TL_\beta\left\|\beta-\beta^0\right\|\right\}.\end{aligned}$$

By choosing $a=\alpha_{g_i^0 t}$, $b=\alpha^0_{g_i^0}$, and $\beta=\beta^0$ in (44), I have, for all (β,α) and all i,

$$\begin{aligned}W_{ig}(\beta,\alpha)\leq{}&\mathbf{1}\left\{g_i^0\neq g\right\}\\ &\times\mathbf{1}\left\{\sum_{t=1}^{T}\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta^0+\alpha^0_{g_i^0 t}\right)\right)\right)\leq\sum_{t=1}^{T}\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+\alpha_{gt}\right)\right)\right)\right.\\ &\left.+TL_\beta\left\|\beta-\beta^0\right\|+TL_\alpha\left\|\alpha^0_{g_i^0}-\alpha_{g_i^0}\right\|\right\},\end{aligned}$$

where I used the norm inequality $\left\|u\right\|_1 \leq \sqrt{T}\left\|u\right\| \leq T\left\|u\right\|$ for all $u \in \mathbb{R}^T$, $T \in \mathbb{N}^*$, where $\left\|\cdot\right\|_1$ is the ℓ^1-norm. Next, a second-order Taylor expansion of $z \mapsto \ln \Psi(z)$ at $Q_{it}Z_{it}$ around $Q_{it}Z_{it}^0$ combined with (A.3), yields

$$
\begin{aligned}
W_{ig}(\beta,\alpha) \leq\ & \mathbf{1}\left\{g_i^0 \neq g\right\} \\
& \times \mathbf{1}\left\{0 \leq \sum_{t=1}^{T} \frac{Y_{it} - \Psi\left(Z_{it}^0\right)}{\Psi\left(Z_{it}^0\right)\left(1-\Psi\left(Z_{it}^0\right)\right)} \Psi'\left(Z_{it}^0\right)\left(X_{it}'\left(\beta-\beta^0\right) + \alpha_{gt} - \alpha_{g_i^0 t}^0\right)\right. \\
& \left. - \frac{b_{\min}}{2}\left(X_{it}'\left(\beta-\beta^0\right) + \alpha_{gt} - \alpha_{g_i^0 t}^0\right)^2 + TL_\beta\left\|\beta-\beta^0\right\| + TL_\alpha\left\|\alpha_{g_i^0}^0 - \alpha_{g_i^0}\right\|\right\} \\
\leq\ & \max_{\widetilde{g}\neq g} \mathbf{1}\left\{0 \leq \sum_{t=1}^{T}\left[\frac{Y_{it} - \Psi\left(Z_{it}^0\right)}{\Psi\left(Z_{it}^0\right)\left(1-\Psi\left(Z_{it}^0\right)\right)} \Psi'\left(Z_{it}^0\right)\left(X_{it}'\left(\beta-\beta^0\right) + \alpha_{gt} - \alpha_{\widetilde{g}t}^0\right)\right.\right. \\
& \left.\left. - \frac{b_{\min}}{2}\left(X_{it}'\left(\beta-\beta^0\right) + \alpha_{gt} - \alpha_{\widetilde{g}t}^0\right)^2\right] + TL_\beta\left\|\beta-\beta^0\right\| + TL_\alpha\left\|\alpha_{\widetilde{g}}^0 - \alpha_{\widetilde{g}}\right\|\right\},
\end{aligned}
$$

Now, let define $V_{it} = \frac{Y_{it}-\Psi\left(Z_{it}^0\right)}{\Psi\left(Z_{it}^0\right)\left(1-\Psi\left(Z_{it}^0\right)\right)}\Psi'\left(Z_{it}^0\right)$, and

$$
\begin{aligned}
A_T = & \left|\sum_{t=1}^{T}\left[V_{it}\left(X_{it}'\left(\beta-\beta^0\right) + \alpha_{gt} - \alpha_{\widetilde{g}t}^0\right) - \frac{b_{\min}}{2}\left(X_{it}'\left(\beta-\beta^0\right) + \alpha_{gt} - \alpha_{\widetilde{g}t}^0\right)^2\right] + TL_\beta\left\|\beta-\beta^0\right\|\right. \\
& \left. + TL_\alpha\left\|\alpha_{\widetilde{g}}^0 - \alpha_{\widetilde{g}}\right\| - \sum_{t=1}^{T} V_{it}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) + \frac{b_{\min}}{2}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2\right|.
\end{aligned}
$$

As I have

$$
\begin{aligned}
A_T \leq & \left|\sum_{t=1}^{T} V_{it}X_{it}'\left(\beta-\beta^0\right)\right| + \left|\sum_{t=1}^{T} V_{it}\left(\alpha_{gt} - \alpha_{\widetilde{g}t}^0\right) - \sum_{t=1}^{T} V_{it}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)\right| + \frac{b_{\min}}{2}\left|\sum_{t=1}^{T} X_{it}'\left(\beta-\beta^0\right)\right| \\
& + b_{\min}\left|\sum_{t=1}^{T} X_{it}'\left(\beta-\beta^0\right)\left(\alpha_{gt} - \alpha_{\widetilde{g}t}^0\right)\right| + \frac{b_{\min}}{2}\left|\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{gt}\right)\left(\alpha_{gt}^0 - 2\alpha_{\widetilde{g}t}^0\right)\right| \\
& + TL_\beta\left\|\beta-\beta^0\right\| + TL_\alpha\left\|\alpha_{\widetilde{g}}^0 - \alpha_{\widetilde{g}}\right\|,
\end{aligned}
$$

it is easy to show using the CS inequality that, for $(\beta,\alpha) \in \mathcal{N}_\eta$,

$$
\begin{aligned}
A_T \leq\ & T\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} V_{it}^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|X_{it}\right\|^2\right)^{1/2} + TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T} V_{it}^2\right)^{1/2} \\
& + b_{\min}\left(\frac{1}{2} + 2\sup_{\alpha\in\mathcal{A}}|\alpha|\right)\sqrt{\eta}\sum_{t=1}^{T}\left\|X_{it}\right\| \\
& + T\sqrt{\eta}\frac{3b_{\min}}{2}\sup_{\alpha\in\mathcal{A}}\left\|\alpha\right\| + T\sqrt{\eta}\left(L_\beta + L_\alpha\right) \\
\leq\ & T\sqrt{\eta}\left[(c_1 \vee c_2) \times (M + C_1) + b_{\min}C_2M + C_3 + L_\beta + L_\alpha\right],
\end{aligned}
$$

where C_1, C_2, C_3,

$$c_1 := \sup_{(\beta,\alpha,g,x)\in\mathcal{B}\times\mathcal{A}^{\mathcal{G}^0 T}\times\mathcal{G}^0\times\cup_{t=1,\ldots,i=1,\ldots}\mathrm{Supp}(X_{it})} \Psi'(Z_{it})/\Psi(Z_{it}),$$
$$c_2 := \sup_{(\beta,\alpha,g,x)\in\mathcal{B}\times\mathcal{A}^{\mathcal{G}^0 T}\times\mathcal{G}^0\times\cup_{t=1,\ldots,i=1,\ldots}\mathrm{Supp}(X_{it})} \Psi'(Z_{it})/(1-\Psi(Z_{it})),$$

are positive constants, independent of η and T. I thus obtain that

$$W_{ig}(\beta,\alpha) \leq \max_{\widetilde{g}\neq g} \mathbf{1}\left\{\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{b_{\min}}{2}\sum_{t=1}^{T}\left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \right.$$
$$\left. + T\sqrt{\eta}\left[(c_1 \vee c_2)\times(M + C_1) + b_{\min}C_2 M + C_3 + L_\beta + L_\alpha\right]\right\}.$$

Noting that the right-hand side of this inequality does not depend on (β,α), it follows that $\sup_{(\beta,\alpha)\in\mathcal{N}_\eta} W_{ig}(\beta,\alpha) \leq \overline{W}_{ig}$, where

$$\overline{W}_{ig} = \max_{\widetilde{g}\neq g} \mathbf{1}\left\{\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{b_{\min}}{2}\sum_{t=1}^{T}\left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \right. \tag{45}$$
$$\left. + T\sqrt{\eta}\left[(c_1 \vee c_2)\times(M + C_1) + b_{\min}C_2 M + C_3 + L_\beta + L_\alpha\right]\right\}. \tag{46}$$

As a result,

$$\sup_{(\beta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i(\beta,\alpha)\neq g^0_i\right\} \leq \frac{1}{N}\sum_{i=1}^{N}\sum_{g\in\mathcal{G}^0}\overline{W}_{ig}. \tag{47}$$

I have, using standard probability algebra and for all $g\in\mathcal{G}^0$,

$$\begin{aligned}
\Pr\left(\overline{W}_{ig}=1\right) &\leq \sum_{\widetilde{g}\neq g}\Pr\Bigg(\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{b_{\min}}{2}\sum_{t=1}^{T}\left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \\
&\qquad + T\sqrt{\eta}\left[(c_1 \vee c_2)\times(M + C_1) + b_{\min}C_2 M + C_3 + L_\beta + L_\alpha\right]\Bigg) \\
&\leq \sum_{\widetilde{g}\neq g}\Bigg\{\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \leq \frac{c_{g,\widetilde{g}}}{2}\right) \\
&\qquad + \Pr\Bigg(\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -T\frac{c_{g,\widetilde{g}}b_{\min}}{4} \\
&\qquad + T\sqrt{\eta}\left[(c_1 \vee c_2)\times(M + C_1) + b_{\min}C_2 M + C_3 + L_\beta + L_\alpha\right]\Bigg)\Bigg\}.
\end{aligned} \tag{48}$$

To end the proof, let $\mathcal{F}_t = \sigma\left(\left\{\mathbf{X}^{(t)}_{-}, \boldsymbol{\varepsilon}^{(t)}_{-}, \gamma^0, \alpha^0\right\}\right)$ denote the σ-field generated by $\mathbf{X}^{(t)}_{-}, \boldsymbol{\varepsilon}^{(t)}_{-}, \gamma^0$, and α^0 and set $S_{it} = \sum_{s=1}^{t} V_{is}\left(\alpha^0_{\widetilde{g}s} - \alpha^0_{gs}\right)$. Then, $\{(S_{it},\mathcal{F}_t), 1\leq t\leq T\}$ is a martingale under

Assumptions 6(a) and 6(b) since

$$\begin{aligned}
&\mathbb{E}\left(\sum_{s=1}^{t} V_{is}\left(\alpha^0_{\widetilde{g}s}-\alpha^0_{gs}\right) \mid \mathcal{F}_{t-1}\right)\\
&=\sum_{s=1}^{t-1} V_{is}\left(\alpha^0_{\widetilde{g}s}-\alpha^0_{gs}\right)+\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)\mathbb{E}\left(\frac{Y_{it}-\Psi\left(Z^0_{it}\right)}{\Psi\left(Z^0_{it}\right)\left(1-\Psi\left(Z^0_{it}\right)\right)}\Psi'\left(Z^0_{it}\right) \mid \mathcal{F}_{t-1}\right)\\
&=\sum_{s=1}^{t-1} V_{is}\left(\alpha^0_{\widetilde{g}s}-\alpha^0_{gs}\right)+\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)\mathbb{E}\left(\mathbb{E}\left(\frac{Y_{it}-\Psi\left(Z^0_{it}\right)}{\Psi\left(Z^0_{it}\right)\left(1-\Psi\left(Z^0_{it}\right)\right)}\Psi'\left(Z^0_{it}\right) \mid \mathcal{F}_{t-1},\sigma\left(\mathbf{X}^{(t)}_{-}\right)\right) \mid \mathcal{F}_{t-1}\right)\\
&=\sum_{s=1}^{t-1} V_{is}\left(\alpha^0_{\widetilde{g}s}-\alpha^0_{gs}\right),
\end{aligned}$$

where the last equality follows from independence of ε_t and $(\mathbf{X}^{(t)}_{-},\boldsymbol{\varepsilon}^{(t-1)}_{-},\gamma^0,\alpha^0)$ and

$$\mathbb{E}\left(Y_{it} \mid X_{i1},...,X_{it},\alpha^0,\gamma^0\right)-\Psi\left(Z^0_{it}\right)=0.$$

By Assumption 7(b), for all $i\in\{1,...,N\}$, $\left\{V_{it}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right):t\right\}$ is such that $\left|V_{it}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)\right|\leq(\widetilde{c}_1\vee\widetilde{c}_2)<\infty$, where the positive constants $\widetilde{c}_j=2c_j\sup_{\alpha\in\mathcal{A}}|\alpha|>0$, for $j\in\{1,2\}$, do not depend on (i,t). Let $a>0$. By Lemma 5, there exist positive constants A and B, independent from (i,t), such that for all $z>a/\sqrt{T}$,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)\right|\geq z\right)\leq A\exp\left(-B(z^2T)^{1/3}\right). \tag{49}$$

I now bound the two terms on the right-hand side of (48).

- By applying Lemma 4, and conducting the same reasoning as in the first bullet point page 1176 in Bonhomme and Manresa (2015), under Assumptions 7(a) and 8(b)-(c), for all $\delta>0$ and as T tends to infinity,

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{gt}-\alpha^0_{\widetilde{g}t}\right)^2\leq\frac{c_{g,\widetilde{g}}b_{\min}}{2}\right)=o(T^{-\delta}).$$

- Lastly, to bound the second term on the right-hand side of (48), I denote as $\underline{c}$ the minimum of $c_{g,\widetilde{g}}$ over all $g\neq\widetilde{g}$ and I take

$$\eta\leq\left(\frac{\underline{c}}{8[(c_1\vee c_2)\times(M+C_1)+b_{\min}C_2M+C_3+L_\beta+L_\alpha]}\right)^2. \tag{50}$$

  Note that this upper bound on η does not depend on T. Taking η satisfying (50) yields, for all $\widetilde{g} \neq g$,

$$\Pr\left(\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -T\frac{c_{g,\widetilde{g}}b_{\min}}{4} + T\sqrt{\eta}\left[(c_1 \vee c_2) \times (M + C_1) + b_{\min}C_2M + C_3 + L_\beta + L_\alpha\right]\right)$$
$$\leq \Pr\left(\frac{1}{T}\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{c_{g,\widetilde{g}}}{8}\right).$$

Lastly, by applying (49) with $z = \frac{c_{g,\widetilde{g}}}{8}$, for T sufficiently large, I obtain

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T} V_{it}\left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{c_{g,\widetilde{g}}}{8}\right) = O(\exp(-C_3T^{1/3})) = o(T^{-\delta}), \tag{51}$$

for all $\delta > 0$, and for some constant C_3 that does not depend on i, T, and g.

Combining results, I thus obtain, using (48), that for η satisfying (50) and for all $\delta > 0$,

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{g\in\mathcal{G}^0}\Pr\left(\overline{W}_{ig} = 1\right) \leq \left|\mathcal{G}^0\right|\left(\left|\mathcal{G}^0\right| - 1\right)\left[o(T^{-\delta}) + o(T^{-\delta})\right] = o(T^{-\delta}). \tag{52}$$

To complete the proof of Lemma 6, note that, for η that satisfies (50), I have, for all $\delta > 0$ and all $\varepsilon > 0$,

$$\Pr\left(\sup_{(\beta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i(\beta,\alpha) \neq g^0_i\right\} > \varepsilon T^{-\delta}\right) \leq \Pr\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g\in\mathcal{G}^0}\overline{W}_{ig} > \varepsilon T^{-\delta}\right)$$
$$\leq \frac{\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g\in\mathcal{G}^0}\overline{W}_{ig}\right)}{\varepsilon T^{-\delta}} = o(1),$$

where I have used (47), the Markov inequality, and (52), respectively. This ends the proof of Lemma 6. □

I am now in position to prove the three parts of the following asymptotic equivalence result.

Lemma 7 (Asymptotic Equivalence) *Let Assumptions 6, 7, and 8 hold. Then, for all $\delta > 0$ and as N and T tend to infinity*

$$\Pr\left(\sup_{i\in\{1,\ldots,N\}}\left|\widehat{g}_i - g^0_i\right| > 0\right) = o(1) + o(NT^{-\delta}), \tag{53}$$

and

$$\widehat{\beta} = \widetilde{\beta} + o_p(T^{-\delta}), \tag{54}$$

and

$$\widehat{\alpha}_{gt} = \widetilde{\alpha}_{gt} + o_p(T^{-\delta}) \text{ for all } g, t. \tag{55}$$

Proof of Lemma 7: The proof closely follows pages 1178-1180 in Bonhomme and Manresa (2015).

#1. Properties of $\widehat{\beta}$. Define

$$\widehat{Q}(\beta,\alpha) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{\widehat{g}_i(\beta,\alpha)t}\right)\right)\right), \tag{56}$$

$$\widetilde{Q}(\beta,\alpha) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{g_i^0 t}\right)\right)\right). \tag{57}$$

Notice that $\widehat{Q}(\cdot)$ is minimized at $(\widehat{\beta},\widehat{\alpha})$ and $\widetilde{Q}(\cdot)$ is minimized at $(\widetilde{\beta},\widetilde{\alpha})$. Let $\eta > 0$ be small enough such that the conclusion of Lemma 6 holds. Using Assumptions 7(a) and 7(b), it is then easy to see that, for all $\delta > 0$,

$$\sup_{(\beta,\alpha)\in\mathcal{N}_\eta}\left|\widehat{Q}(\beta,\alpha) - \widetilde{Q}(\beta,\alpha)\right| = o_p(T^{-\delta}). \tag{58}$$

By consistency of $\widehat{\beta}$ (Theorem 2) and $\widehat{\alpha}$ (Lemma 3), and because $\widetilde{\beta}$ and $\widetilde{\alpha}$ are also consistent under the conditions of Theorem 2, I have, as N and T tend to infinity,

$$\Pr\left(\left(\widehat{\beta},\widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) \to 0, \tag{59}$$

$$\Pr\left(\left(\widetilde{\beta},\widetilde{\alpha}\right) \notin \mathcal{N}_\eta\right) \to 0. \tag{60}$$

Then, the same arguments as those appearing between (B-14) and (B-17) in page 1179 in Bonhomme and Manresa (2015) can be used to show that eq. (58)-(60) imply

$$\widetilde{Q}(\widehat{\beta},\widehat{\alpha}) - \widetilde{Q}(\widetilde{\beta},\widetilde{\alpha}) = o_p(T^{-\delta}). \tag{61}$$

Now, using that $(\widetilde{\beta},\widetilde{\alpha})$ minimizes the twicely continuously differentiable function $\widetilde{Q}(\cdot)$, I obtain under Assumption 6(b)

$$\begin{aligned}
\widetilde{Q}(\widehat{\beta},\widehat{\alpha}) - \widetilde{Q}(\widetilde{\beta},\widetilde{\alpha}) &\geq \frac{b_{\min}}{2}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}'\left(\widetilde{\beta}-\widehat{\beta}\right) + \widetilde{\alpha}_{g_i^0 t} - \widehat{\alpha}_{g_i^0 t}\right)^2, \\
&\geq \frac{b_{\min}}{2}\left(\widetilde{\beta}-\widehat{\beta}\right)'\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\overline{X}_{g_i^0 t}\right)\left(X_{it}-\overline{X}_{g_i^0 t}\right)'\right)\left(\widetilde{\beta}-\widehat{\beta}\right) \\
&\geq \frac{\widehat{\rho} b_{\min}}{2}\left\|\widetilde{\beta}-\widehat{\beta}\right\|^2,
\end{aligned}$$

where $\widehat{\rho} \xrightarrow{p} \rho > 0$ as a consequence of Assumption 7(c). Hence, $\widetilde{\beta} - \widehat{\beta} = o_p(T^{-\delta})$ for all $\delta > 0$. This shows (54).

#2. Properties of $\widehat{\alpha}$. The proof is identical to page 1180 in Bonhomme and Manresa (2015).
#3. Properties of $\widehat{g}_i = \widehat{g}_i(\widehat{\beta},\widehat{\alpha})$. The proof is identical to page 1180 in Bonhomme and Manresa (2015).

The proof of Lemma 7 is complete. □

A.4.2 Step 2: Asympotic Properties of the Oracle MLE

By Lemma 7 and Slutsky's lemma, it is sufficient to analyze the limiting distribution of the unfeasible maximum likelihood estimator, $(\widetilde{\beta}, \widetilde{\alpha})$, defined as

$$(\widetilde{\beta}, \widetilde{\alpha}) = \underset{(\beta,\alpha)\in\mathcal{B}\times\mathcal{A}^{G^0 T}}{\arg\min} \widetilde{Q}(\beta, \alpha),$$

where

$$\widetilde{Q}(\beta, \alpha) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g\in\mathcal{G}^0} \mathbf{1}\left\{g_i^0 = g\right\} \times \left[-\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{gt}\right)\right)\right)\right].$$

First, I show

$$\sqrt{NT}\left(\widetilde{\beta} - \beta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta^{-1}\right). \tag{62}$$

Second, I show for all g, t,

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha_{gt}^0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\widetilde{\pi}_{gt}^2}\right), \tag{63}$$

and conclude by Slutsky's lemma.

\# **1. (62) holds.** Under Assumption 9, results in Hahn and Newey (2004) (eq. (3)) and Arellano and Hahn (2007) (in case of multi-dimensional fixed effects of size G^0) ensure

$$\sqrt{NT}\left(\widetilde{\beta} - \beta^0\right) = S_{NT} + \sqrt{\frac{T}{N}}B + O_p\left(\sqrt{\frac{T}{N^3}}\right),$$

for some deterministic $B \in \mathbb{R}^{p\times p}$ and $S_{NT} \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta^{-1}\right)$. The result then follows from $T = o(N)$.

#2. (63) holds. Let $(g, t) \in \mathcal{G}^0 \times \mathbb{N}^*$. For all $\beta \in \mathcal{B}$, define the optimal $\widetilde{\alpha}_{gt}(\beta)$ as

$$\widetilde{\alpha}_{gt}(\beta) = \underset{\alpha\in\mathcal{A}}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} -\mathbf{1}\left\{g_i^0 = g\right\} \times \ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha\right)\right)\right).$$

The first-order optimality condition for $\widetilde{\alpha}_{gt}(\beta)$ writes

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{g_i^0 = g\right\} Q_{it}\left(\ln\Psi\right)'\left(Q_{it}\left(X_{it}'\beta + \widetilde{\alpha}_{gt}(\beta)\right)\right) = 0. \tag{64}$$

Differentiating eq. (64) with respect to β yields

$$\frac{\mathrm{d}\widetilde{\alpha}_{gt}(\beta)}{\mathrm{d}\beta} = -\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0 = g\right\}\left(\ln\Psi_{it}\right)''\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0 = g\right\}\left(\ln\Psi_{it}\right)'' X_{jt}\right), \tag{65}$$

where $(\ln\Psi_{it})'' := (\ln\Psi)''\left(Q_{it}\left(X_{it}'\beta + \widetilde{\alpha}_{g_i^0 t}(\beta)\right)\right)$. By Taylor's theory, eq. (65) and Assumptions 7(a)-

(b) imply that there exists $C > 0$ such that, almost surely,

$$\sup_{\beta,\beta' \in \mathcal{B}} |\widetilde{\alpha}_{gt}(\beta) - \widetilde{\alpha}_{gt}(\beta')| \leq C \left\| \beta - \beta' \right\|. \tag{66}$$

Deduce that

$$\begin{aligned}
\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha_{gt}^0\right) &= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0\right) + \sqrt{N}\left(\widetilde{\alpha}_{gt}(\widetilde{\beta}) - \widetilde{\alpha}_{gt}(\beta^0)\right) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0\right) + O_p\left(\sqrt{N}\left\|\widetilde{\beta} - \beta^0\right\|\right) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0\right) + O_p(1/\sqrt{T}) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0\right) + o_p(1),
\end{aligned} \tag{67}$$

where the second and third equality use eq. (66) and (62) respectively. Now, by expanding each summand in eq. (64) at $X_{it}'\beta^0 + \widetilde{\alpha}_{gt}(\beta^0)$ around Z_{it}^0, Taylor's theory ensures again that there exists $Z_{it}^* \in \mathcal{Z}$ such that

$$\widetilde{\alpha}_{gt}(\beta^0) = \alpha_{gt}^0 - \left(\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}(-\ln\Psi)''\left(Q_{it}Z_{it}^*\right)\right)^{-1}\left(\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}Q_{it}(-\ln\Psi)'\left(Q_{it}Z_{it}^0\right)\right). \tag{68}$$

Equation (68) yields

$$\begin{aligned}
&\sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0\right) \\
&= -\left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}(-\ln\Psi)''\left(Q_{it}Z_{it}^*\right)\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}Q_{it}(-\ln\Psi)'\left(Q_{it}Z_{it}^0\right)\right) \\
&= \left(\widetilde{\pi}_{gt}^{-1} + o_p(1)\right)\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}Q_{it}(\ln\Psi)'\left(Q_{it}Z_{it}^0\right)\right) \\
&\xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\widetilde{\pi}_{gt}^2}\right),
\end{aligned}$$

where the second equality follows from $\sup_{i=1,\ldots,N}|Z_{it}^* - Z_{it}^0| = o_p(1)$ (it is easy to prove that $\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0 = o_p(1)$ using (68), Assumptions 6(b), 7(a)-(b), and 9(e)) and Assumption 9(c), and the last convergence follows by Assumption 9(e). Given (67), (63) follows by Slutsky's lemma.

#3. Conclusion. Let $\delta > 0$. By Lemma 7,

$$\begin{aligned}
\sqrt{NT}\left(\widehat{\beta} - \beta^0\right) &= \sqrt{NT}\left(\widetilde{\beta} - \beta^0\right) + \sqrt{NT}\left(\widehat{\beta} - \widetilde{\beta}\right) \\
&= \sqrt{NT}\left(\widetilde{\beta} - \beta^0\right) + o_p\left(\sqrt{N}T^{1-\delta}\right),
\end{aligned} \tag{69}$$

and, for all $g \in \mathcal{G}^0$, all $t \in \mathbb{N}^*$,

$$\begin{aligned}\sqrt{N}\left(\widehat{\alpha}_{gt}-\alpha^0_{gt}\right) &= \sqrt{N}\left(\widetilde{\alpha}_{gt}-\alpha^0_{gt}\right)+\sqrt{N}\left(\widehat{\alpha}_{gt}-\widetilde{\alpha}_{gt}\right)\\ &= \sqrt{N}\left(\widetilde{\alpha}_{gt}-\alpha^0_{gt}\right)+o_p\left(\sqrt{N}T^{-\delta}\right).\end{aligned} \tag{70}$$

Since (69) and (70) hold for all $\delta > 0$, and there exists $\nu > 0$ such that $N/T^{\nu} \to 0$, as N and T tend to infinity, I obtain

$$\begin{aligned}\sqrt{NT}\left(\widehat{\beta}-\beta^0\right) &= \sqrt{NT}\left(\widetilde{\beta}-\beta^0\right)+o_p(1),\\ \sqrt{N}\left(\widehat{\alpha}_{gt}-\alpha^0_{gt}\right) &= \sqrt{N}\left(\widetilde{\alpha}_{gt}-\alpha^0_{gt}\right)+o_p(1).\end{aligned}$$

This result, combined with (62), (63), and Slustky's lemma yields (20) and (21).

B Extensions

B.1 Cluster-Specific Slopes and Time-Specific Effects

In this section, I consider the following extension of model (2): for all $(i,t) \in \mathcal{N} \times \mathcal{T}$,

$$\Pr\left(Y_{it}=y \mid X_{i1},...,X_{it},\alpha^0_{g_i^0 t},\beta^0_{g_i^0},g_i^0,\zeta_t^0\right)=h^0\left(y,X_{it}'\beta^0_{g_i^0}+\alpha^0_{g_i^0 t}+\zeta_t^0\right), \tag{71}$$

where $h^0 \in \mathcal{H}$, $\left\|\beta_1^0\right\| = 1$ and $\alpha^0_{11} = \zeta^0_1 = 0$ are normalizations. Absent of correlation between the groups and if groups were known, I could just run separate analysis of each panel data $\{(i,t) \in \mathcal{N} \times \mathcal{T} : g_i^0 = g\}_{g\in\mathcal{G}^0}$. Here, the difficulty arises from the assumption that the group membership variables g_i^0 are unknown to the econometrician. Let $\beta^0 := \{\beta^0_g : g\}$. I first adapt Assumption 1:

Assumption 11 (Random sampling)

(a) $(Y_i', X_i', g_i^0)'$ *is i.i.d. across* $i \in \mathcal{N}$ *conditional on* α^0, β^0, λ^0, μ^0.

(b) For all $i \in \mathcal{N}$: $\{(Y_{it}, X_{it}', \alpha^0_{g_i^0 t}, \zeta_t^0)'\}_{t\geq 2}$ *is a strictly stationary strong mixing process with mixing coefficients* $\tau_i(\cdot)$ *conditional on* $g_i^0, \mu^0_{g_i^0}, \xi_i^0, \beta^0_{g_i^0}$. *Let* $\tau(\cdot) = \sup_i \tau_i(\cdot)$ *satisfy* $\tau(l) \leq Cm^l$ *with* $C > 0$, *and* $m \in (0,1)$.

(c) For all $t \in \mathcal{T}$: $Y_{1t} \mid X_{1t}, g_1, \alpha^0, \beta^0, \lambda^0, \mu^0, \xi^0 \overset{d}{=} Y_{1t} \mid X_{1t}, g_1^0, \alpha^0_{g_1^0 t}, \beta^0_{g_i^0}$.

Assumption 12 (Latent clustering)

(a) There exist known $\mathcal{X}^0 \subset \mathcal{X}$, $y \in \mathcal{Y}$, *and functional* ϕ *such that, for all fixed* $(i,j) \in \mathcal{N}^2$, *letting* $\rho_i(x) : \mathcal{X}^0 \ni x \mapsto \Pr\left(Y_{i2} = y \mid X_{i2} = x, \beta^0_{g_i^0}, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right)$, $\phi(\rho_i, \rho_j) = \mathbb{1}\{g_i^0 = g_j^0\}$.

(b) For all $g \in \mathcal{G}^0$, *almost surely* $\Pr(g_1^0 = g \mid \alpha^0, \beta^0, \lambda^0, \mu^0, \xi^0) > 0$.

Assumption 13 (Regularity and smoothness)

(a) *Conditional on* $g_i^0, \mu^0_{g_i^0}, \xi_i^0, \beta^0_{g_i^0}$, X_{i2} *admits a uniformly continuous density function* $f_{X_{i2}|g_i^0,\mu^0_{g_i^0},\xi_i^0,\beta^0_{g_i^0}}$ *such that* $\inf_{x\in\mathcal{X}^0} f_{X_{i2}|g_i^0,\mu^0_{g_i^0},\xi_i^0,\beta^0_{g_i^0}}(x) \geq \delta > 0$ *and* $\sup_{x\in\mathcal{X}^0} f_{X_{i2}|g_i^0,\mu^0_{g_i^0},\xi_i^0,\beta^0_{g_i^0}}(x) < \infty$.

(b) *Almost surely,* $\mathbb{E}(\left\|X_{12}\right\|^2 \mid g_1^0, \alpha^0, \beta^0, \lambda^0, \mu^0)$ *is finite and* $\mathbb{E}(X_{12}X_{12}' \mid g_1^0, \alpha^0, \beta^0, \lambda^0, \mu^0)$ *is non-singular.*

(c) *For all* $g \in \mathcal{G}^0$*:* $\sum_{y\in\mathcal{Y}} yh^0(y,\cdot)$ *is differentiable on* $\mathbb{R}$ *and not constant on the support of* $X_{it}'\beta^0_{g_i^0} + \alpha^0_{g_i^0 t}$.

Assumption 14 (Monotonicity) *There exists* $y \in \mathcal{Y}$ *such that* $h^0(y, v)$ *is strictly monotonic in* v.

Assumption 15 (Compensating variations)

(a) *For all fixed* $(g, t, \tilde{t})$*, all* $x_1 \in \mathcal{X}$*, there exists* $x_2 \in \mathcal{X}$ *such that*

$$\alpha^0_{gt} + x_1'\beta^0_g + \zeta^0_t = \alpha^0_{g\tilde{t}} + x_2'\beta^0_g + \zeta^0_{\tilde{t}}. \tag{72}$$

(b) *For all fixed* $(g, \tilde{g}, t)$*, all* $x_3 \in \mathcal{X}$*, there exists* $x_4 \in \mathcal{X}$ *such that*

$$\alpha^0_{gt} + x_3'\beta^0_g + \zeta^0_t = \alpha^0_{\tilde{g}t} + x_4'\beta^0_{\tilde{g}} + \zeta^0_t. \tag{73}$$

Theorem 4 (Identification) *Let Assumptions 11, 12 and 13(a) hold, and let* N *and* T *diverge jointly to infinity.*

1. $\{W_N^0 : N \in \mathbb{N}^*\}$ *and* G^0 *are identified.*
2. *If Assumptions 13(b)-15 further hold, then*
   - β^0 *is identified.*
   - $\zeta^0_t + \alpha^0_{gt}$ *is identified for all* $(g,t) \in \mathcal{G}^0 \times \mathbb{N}^*$.

Proof of Theorem 4: The proofs of Part 1 and identification of β^0 are identical to the corresponding parts of the proof of Theorem 1, up to running nonparametric regressions for all $g \in \mathcal{G}^0$ to identify β^0_g. Next, Assumption 15(b) ensures that, for all $(g, \tilde{g}, t)$, I can identify $(x_1, x_2) \in \mathcal{X}^2$, such that for some $y \in \mathcal{Y}$,

$$h^0\left(y, x_1'\beta^0_g + \alpha^0_{gt} + \zeta^0_t\right) = h^0\left(y, x_2'\beta^0_{\tilde{g}} + \alpha^0_{\tilde{g}t} + \zeta^0_t\right).$$

By inverting $h^0(y,\cdot)$, I obtain $\alpha^0_{gt} - \alpha^0_{\tilde{g}t} = x_1'\beta^0_{\tilde{g}} - x_2'\beta^0_g$. Since the right-hand side is identified, $\alpha^0_{gt} - \alpha^0_{\tilde{g}t}$ is identified for all $(g, \tilde{g}, t)$. In particular, $(\alpha^0_{g1})_{g\in\mathcal{G}^0}$ is identified. Now, suppose that $G^0 \geq 2$. By Assumption 15(a), for all $(g, t, \tilde{t})$, I can identify $(x_3, x_4) \in \mathcal{X}^2$ such that, for some $y \in \mathcal{Y}$,

$$h^0\left(y, x_3'\beta^0_g + \alpha^0_{gt} + \zeta^0_t\right) = h^0\left(y, x_4'\beta^0_g + \alpha^0_{g\tilde{t}} + \zeta^0_{\tilde{t}}\right). \tag{74}$$

By inverting $h^0(y,\cdot)$ again, eq. (74) yields

$$\zeta_t^0 - \zeta_{\tilde{t}}^0 = \alpha_{g\tilde{t}}^0 - \alpha_{gt}^0 + (x_4 - x_3)'\beta_g^0. \tag{75}$$

Because $\zeta_1^0 = \alpha_{11}^0 = 0$, $\zeta_t^0 + \alpha_{1t}^0$ and $\zeta_t^0 + \alpha_{gt}^0 = \zeta_t^0 + \alpha_{1t}^0 + \alpha_{gt}^0 - \alpha_{1t}^0$ are identified for all (g,t).

B.2 Group and Time-Specific Link Functions

Consider the general model:

$$\Pr\left(Y_{it} = y \mid X_i^t, g_i^0\right) = h_t^0\left(y, X_{it}'\theta^0, g_i^0\right), \quad i = 1, ..., N, t = 1, ..., T. \tag{76}$$

Under an adaptation of Assumption 2, the same analysis can be conducted to identify g_i^0 and $(h_t^0)_{t\geq 1}$ up to group relabeling, and θ^0 up to scale.

B.3 Grouping Time Periods

Consider a model in which time effects are also grouped: there exists $(g_i^0, k_t^0) \in \{1, ..., G^0\} \times \{1, ..., K^0\}$ such that

$$\Pr\left(Y_{it} = y \mid X_i^t, \alpha_{g_i^0 k_t^0}^0, g_i^0, k_t^0\right) = h^0\left(y, X_{it}'\theta^0 + \alpha_{g_i^0 k_t^0}^0\right), \quad i = 1, ..., N, t = 1, ..., T \tag{77}$$

When $\mathcal{N} = \mathcal{T}$, this gives rise to a so-called Holland, Laskey, and Leinhardt (1983)'s stochastic block model on latent variables. Methods developed in the present paper and in Mugnier (2022) can be used to obtain identification results for nonlinear multiplicative models in cases where $G^0 = K^0$ and under symmetry ($\alpha_{g\tilde{g}}^0 = \alpha_{\tilde{g}g}^0$ almost surely).

B.4 NGFE Large Sample Theory for Poisson Count Models

Theorem 2 can be generalized to NGFE models satisfying certain moment and concavity/regularity conditions on the series of partial derivatives of $(\beta, \pi) \mapsto \ln h^0(Y_{it}, X_{it}'\beta + \pi) \equiv \ell_{it}(\beta, \pi)$.

Assumption 16

(a) *Smoothness and moments: $(\beta, \pi) \mapsto \ell_{it}(\beta, \pi)$ is three times continuously differentiable almost surely. The partial derivatives of $\ell_{it}(\beta, \pi)$ with respect to the elements of (β, π) up to the second order are bounded in absolute value uniformly over $(\beta, \pi) \in \mathcal{B} \times \mathcal{A}$ by a function $M(Y_{it}, X_{it}) > 0$ almost surely, and*

$$\max_{i,t} E\left[M(Y_{it}, X_{it})^4 \mid \mathbf{X}^{(t)}, \alpha_{g_i^0 t}^0\right]$$

is almost surely uniformly bounded over N, T.

(b) *Strict concavity: for all* N, T, $\frac{\partial^2 \ell_{it}(\beta,\pi)}{\partial \pi^2} < 0$ *almost surely for all* $(\beta, \pi) \in \mathbb{R}^{p+1}$.

In particular, Assumption 16(b) is verified by the Poisson count model (3).

Theorem 5 (Consistency in General Nonlinear Models) *Let Assumptions 7 and 16 hold. Then, as* N *and* T *tend to infinity:*

1. $\widehat{\beta} \xrightarrow{p} \beta^0$, *and*

2. $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2 \xrightarrow{p} 0.$

The proof is available upon request.

Under the existence of a moment generating function for the score on a small interval around zero, the concentration inequalities and most of the arguments in the proof of Theorem 3 could still be applied to obtain asymptotic normality. A technical difficulty here is that Y_{it} is not bounded anymore so that uniform Lipschitz continuity in eq. (44) and (43) does not hold anymore. I only state the result without proof for the Poisson count model. I denote as $\widetilde{X}_{gt}$ the projection of X_{it} on the space spanned by the cluster membership variable under a metric weighted by $\exp(Z^0_{it})$,

$$\widetilde{X}_{gt} = \left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}\exp(Z^0_{it})\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}\exp(Z^0_{it})X_{it}\right),$$

i.e., the weighted mean of X_{it} in cluster $g_i^0 = g$. Also, let define the weighted average

$$\widehat{\pi}_{gt} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}\exp(Z^0_{it}).$$

Consider the following assumption.

Assumption 17

(a) $\{(Y_{it}, X'_{it})' : (i,t)\}$ *are independent conditional on the fixed effects.*

(b) *There exists a positive definite matrix* Σ_β *such that*

$$\Sigma_\beta = \text{plim}_{N,T\to\infty}\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \exp(Z^0_{it})\left[X_{it} - \widetilde{X}_{g_i^0 t}\right]\left[X_{it} - \widetilde{X}_{g_i^0 t}\right]'.$$

(c) *As* N *and* T *tend to infinity,*

$$\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T \left\{\exp(Z^0_{it})\left(X_{it} - \widetilde{X}_{g_i^0 t}\right)\right\}\left\{Y_{it} - \exp(Z^0_{it})\right\} \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta).$$

(d) *For all* (g,t): $\text{plim}_{N\to\infty}\widehat{\pi}_{gt} = \widetilde{\pi}_{gt} > 0$.

(e) *For all* (g,t):

$$\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}E\left(\mathbf{1}\left\{g_i^0=g\right\}\mathbf{1}\left\{g_j^0=g\right\}(Y_{it}-\exp(Z_{it}^0))(Y_{jt}-\exp(Z_{jt}^0))\right)=\omega_{gt}>0.$$

(f) *For all* (g,t), *and as* N *and* T *tend to infinity:*

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0=g\right\}(Y_{it}-\exp(Z_{it}^0))\xrightarrow{d}\mathcal{N}(0,\omega_{gt}).$$

(g) *The true value of* β, β^0, *is in the interior of* $\mathcal{B}$. *For all* T, *the true value of* $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}^0$, *is in the interior of* $\mathcal{A}^{G^0T}$.

Theorem 6 (Asymptotic Distribution in the Poisson Count Model – Conjectured) *Let eq.* (4), *Assumptions 7, 8, and 17 hold, and let* N *and* T *tend to infinity such that* $N/T\to\infty$ *and, for some* $\nu>0$, $N/T^{\nu}\to 0$. *Then:*

$$\sqrt{NT}\left(\widehat{\beta}-\beta^0\right)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_{\beta}^{-1}\right), \tag{78}$$

and, for all (g,t),

$$\sqrt{N}\left(\widehat{\alpha}_{gt}-\alpha_{gt}^0\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\omega_{gt}}{\widetilde{\pi}_{gt}^2}\right), \tag{79}$$

where Σ_{β}, ω_{gt}, *and* $\widetilde{\pi}_g$ *are defined in Assumption 17.*

C Large-N, Large-T Inference

C.1 Binary Choice Model

Assuming independent observations across individual units, the asymptotic variance of $\widehat{\alpha}_{gt}$ for all g,t can be estimated as

$$\mathrm{Var}\left(\widehat{\alpha}_{gt}\right)=\frac{\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i=g\right\}\left((\ln\Psi)'\left(Q_{it}\left(X_{it}'\widehat{\beta}+\widehat{\alpha}_{\widehat{g}_it}\right)\right)\right)^2}{\left(\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i=g\right\}(-\ln\Psi)''\left(Q_{it}\left(X_{it}'\widehat{\beta}+\widehat{\alpha}_{\widehat{g}_it}\right)\right)\right)^2}. \tag{80}$$

Given Theorem 3, an estimate of the asymptotic variance of $\widehat{\beta}$ is

$$\mathrm{Var}\left(\widehat{\beta}\right)=\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(-\ln\Psi)''\left(Q_{it}\left(X_{it}'\widehat{\beta}+\widehat{\alpha}_{\widehat{g}_it}\right)\right)\left[X_{it}-\widehat{\widetilde{X}}_{\widehat{g}_i,t}\right]\left[X_{it}-\widehat{\widetilde{X}}_{\widehat{g}_i,t}\right]'\right)^{-1}, \tag{81}$$

where

$$\widehat{\widehat{X}}_{gt} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ \widehat{g}_i = g \right\} (\ln \Psi)'' \left(Q_{it} \left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) \right) \right)^{-1} \times \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ \widehat{g}_i = g \right\} (\ln \Psi)'' \left(Q_{it} \left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) \right) X_{it} \right).$$

C.2 Poisson Count Model

Assuming independent observations across individual units, the asymptotic variance of $\widehat{\alpha}_{gt}$ for all g, t can be estimated as

$$\operatorname{Var}\left(\widehat{\alpha}_{gt}\right) = \frac{\sum_{i=1}^{N} \mathbf{1} \left\{ \widehat{g}_i = g \right\} \left(Y_{it} - \exp\left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) \right)^2}{\left(\sum_{i=1}^{N} \mathbf{1} \left\{ \widehat{g}_i = g \right\} \exp\left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) \right)^2}. \tag{82}$$

Given Theorem 6, an estimate of the asymptotic variance of $\widehat{\beta}$ is

$$\operatorname{Var}\left(\widehat{\beta}\right) = \left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \exp\left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) \left[X_{it} - \widehat{\widehat{X}}_{\widehat{g}_i, t} \right] \left[X_{it} - \widehat{\widehat{X}}_{\widehat{g}_i, t} \right]' \right)^{-1}, \tag{83}$$

where

$$\widehat{\widehat{X}}_{gt} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ \widehat{g}_i = g \right\} \exp\left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) \right)^{-1} \times \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ \widehat{g}_i = g \right\} \exp\left(X_{it}' \widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t} \right) X_{it} \right).$$

D More Details on Monte Carlo Experiments

To measure classification accuracy, I focus on three metrics inspired from the binary classification and clustering statistical literature, which are invariant to cluster relabeling.[47] The three metrics write

$$\begin{aligned} \text{R} &\equiv \text{Recall rate} := \frac{TP}{TP + FN}, \\ \text{P} &\equiv \text{Precision rate} := \frac{TP}{TP + FP}, \\ \text{RI} &\equiv \text{Rand Index} := \frac{TP + TN}{TP + TN + FP + FN}, \end{aligned}$$

[47] Bonhomme and Manresa (2015) report a "Misclassification Rate" (M) defined as the minimum of $\sum_{i=1}^{N} \left| \widehat{g}_i - g_i^0 \right| / N$ over all possible cluster relabelings for the $\widehat{g}_i$. Beyond the fact that computing MR can be very demanding for large G^0, it is not totally fair since the final labeling of $\widehat{g}_i$ requires knowledge of g_i^0 to be determined.

where

$$
\begin{aligned}
FP \equiv \text{False Positives} &:= \sum_{i<j} \mathbf{1}\left\{\widehat{g}_i = \widehat{g}_j\right\} \mathbf{1}\left\{g_i^0 \neq g_j^0\right\}, \\
TP \equiv \text{True Positives} &:= \sum_{i<j} \mathbf{1}\left\{\widehat{g}_i = \widehat{g}_j\right\} \mathbf{1}\left\{g_i^0 = g_j^0\right\}, \\
FN \equiv \text{False Negatives} &:= \sum_{i<j} \mathbf{1}\left\{\widehat{g}_i \neq \widehat{g}_j\right\} \mathbf{1}\left\{g_i^0 = g_j^0\right\}, \\
TN \equiv \text{True Negatives} &:= \sum_{i<j} \mathbf{1}\left\{\widehat{g}_i \neq \widehat{g}_j\right\} \mathbf{1}\left\{g_i^0 \neq g_j^0\right\}.
\end{aligned}
$$

The Recall rate (R) measures the ability of the NGFE estimator to predict the same group for pairs of individual who truly belong to the same group. The Precision rate (P) measures how precise the pairing prediction is: among all the predicted pairs of individual sharing the same group, what is the proportion of correct ones? The Rand Index (RI) is the proportion of correctly predicted pair (true or false) made by the algorithm.

Initialization of NGFE I use $1,000$ initialization random points $(\theta'_{\text{init}}, \alpha_{11\text{init}}, ..., \alpha_{G^0T\text{init}})'$ such that $\theta_{\text{init}} = v$ where $v \overset{iid}{\sim} \mathcal{N}(0, (1/4)^2)$ and $\alpha_{gt,\text{init}} = \mu_{g,\text{init}} + w$ where $\mu_{g,\text{init}} \overset{iid}{\sim}\text{Unif}[-4, 4]$ and $w \overset{iid}{\sim} \mathcal{N}(0, (1/4)^2)$.

Computation Having large N is not computationally demanding. When T is very large, computation of the NGFE estimate might be demanding. The methods developed in Mugnier (2022) could be adapted. The statistical asymptotic results are confirmed by increasing (N, T) in unreported simulations.

E Tables & Figures

E.1 Monte Carlo Simulations

Table 1: Bias and Root Mean Squared Error of $\widehat{\beta}$ (Static Model)

		NGFE		CMLE		NLTWFE		2STEPGFE		Pooled OLS		LTWFE		GFE	
DGP	G^0	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
1	2	-0.072	0.268	-0.104	0.551	0.217	0.950	-0.252	1.516	-0.407	0.411	-0.790	0.812	-0.798	0.814
	3	-0.089	0.294	0.294	0.637	0.669	1.000	0.355	0.893	-0.363	0.366	-0.724	0.734	-0.853	0.874
	5	-0.022	0.264	0.167	0.538	0.359	0.824	0.104	0.779	-0.369	0.373	-0.766	0.776	-0.784	0.839
2	2	0.106	0.171	0.010	0.161	0.223	0.302	-0.278	0.309	-0.779	0.780	-0.831	0.831	-0.816	0.818
	3	0.236	0.289	0.014	0.160	0.238	0.309	-0.300	0.345	-0.768	0.769	-0.867	0.867	-0.837	0.841
	5	0.601	0.637	-0.004	0.169	0.250	0.332	-0.324	0.358	-0.747	0.747	-0.916	0.916	-0.853	0.860
3	2	0.352	0.385	-0.001	0.169	0.221	0.313	-0.110	0.211	-0.776	0.777	-0.857	0.857	-0.826	0.827
	3	0.432	0.486	-0.002	0.170	0.219	0.308	-0.066	0.192	-0.788	0.789	-0.859	0.859	-0.845	0.846
	5	0.471	0.499	0.011	0.156	0.235	0.309	-0.057	0.186	-0.787	0.788	-0.858	0.858	-0.833	0.836
4	2	0.040	0.151	-0.002	0.152	0.195	0.269	0.085	0.221	-0.789	0.789	-0.783	0.784	-0.788	0.789
	3	0.095	0.159	0.016	0.124	0.223	0.269	0.109	0.213	-0.776	0.776	-0.778	0.779	-0.790	0.792
	5	0.114	0.178	0.018	0.118	0.222	0.266	0.094	0.204	-0.775	0.775	-0.778	0.779	-0.803	0.809

Notes: Static logit model with $\beta = 1$, $N = 90$, and $T = 7$. G^0 = true number of groups. NGFE (resp. 2STEPGFE and GFE) estimates are based on $1,000$ (resp. 100 and 100) initialization points. Results are averaged across 50 Monte Carlo replications.

Table 2: Classification Accuracy and CPU Time (Static Model)

		NGFE					2STEPGFE						GFE				
DGP	G^0	P	R	RI	M	CPU	P	R	RI	M	CPU	$\widehat{G}$	P	R	RI	M	CPU
1	2	0.51	0.87	0.51	0.44	10.62	0.54	0.24	0.51	0.77	10.19	5.38	0.54	0.55	0.54	0.38	29.27
	3	0.35	0.81	0.42	0.57	11.42	0.37	0.24	0.60	0.75	11.34	5.48	0.36	0.38	0.57	0.55	29.63
	5	0.21	0.80	0.35	0.70	14.75	0.24	0.25	0.69	0.71	11.73	5.88	0.24	0.25	0.69	0.63	83.18
2	2	0.56	0.86	0.57	0.36	8.02	0.64	0.45	0.60	0.53	3.57	3.06	0.61	0.61	0.61	0.29	21.95
	3	0.40	0.85	0.49	0.51	8.52	0.57	0.49	0.70	0.44	4.70	3.64	0.46	0.49	0.64	0.42	22.00
	5	0.22	0.87	0.34	0.69	10.15	0.44	0.53	0.77	0.44	5.78	4.44	0.35	0.40	0.74	0.54	20.93

Notes: Static logit model with $\beta = 1$, $N = 90$, and $T = 7$. G^0 = true number of groups, P = Precision rate, R = Recall rate, RI = Rand Index, M = Misclassification Rate = minimum of $\sum_{i=1}^{N} \mathbf{1}\left\{\widehat{g}_i \neq g_i^0\right\}/N$ over all possible cluster relabelings, CPU = CPU time in seconds computed with Python's `time` command `time.perf_counter()`, $\widehat{G}$ = number of groups estimated by 2STEPGFE. NGFE (resp. 2STEPGFE and GFE) estimates are based on $1,000$ (resp. 100 and 100) initialization points. Results are averaged across 50 Monte Carlo replications.

Table 3: Inference for β (Static Model)

		NGFE			CMLE		
DGP	G^0	SE	SD	.95	SE	SD	.95
1	2	0.16	0.26	0.86	0.15	0.54	0.38
	3	0.17	0.28	0.80	0.16	0.56	0.40
	5	0.17	0.26	0.84	0.15	0.51	0.42
2	2	0.12	0.13	0.82	0.06	0.16	0.52
	3	0.12	0.17	0.46	0.07	0.16	0.62
	5	0.14	0.21	0.08	0.08	0.17	0.66
3	2	0.12	0.16	0.22	0.06	0.17	0.52
	3	0.12	0.22	0.18	0.06	0.17	0.52
	5	0.12	0.16	0.04	0.06	0.16	0.56
4	2	0.12	0.15	0.92	0.05	0.15	0.38
	3	0.13	0.13	0.92	0.05	0.12	0.56
	5	0.13	0.14	0.88	0.05	0.12	0.56

Notes: Static logit model with $\beta = 1$, $N = 90$, and $T = 7$. SE reports the median of the estimates of the analytical standard errors based on the large-N, T analytical variance formula (83) across simulations; SD reports the median of the actual standard deviation across simulations; .95 reports the empirical nonrejection probabilities (nominal size 5%) based on the analytical standard errors estimates. Results are averaged across 50 Monte Carlo replications.

Table 4: Bias and Root Mean Squared Error (Dynamic Model)

		NGFE				CMLE				NLTWFE				2STEPGFE			
		Bias		RMSE		Bias		RMSE		Bias		RMSE		Bias		RMSE	
DGP	G^0	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
1	2	-0.026	-0.128	0.229	0.328	-0.663	-0.174	0.689	0.526	-0.702	0.242	0.737	0.965	-0.032	-0.456	0.309	0.666
	3	0.073	-0.144	0.323	0.447	-0.651	0.238	0.676	0.634	-0.684	0.663	0.716	0.995	-0.142	-0.282	0.254	0.745
	5	0.156	-0.279	0.365	0.448	-0.592	0.090	0.629	0.524	-0.606	0.318	0.659	0.826	-0.051	0.158	0.277	0.492
2	2	0.486	0.043	0.630	0.141	-0.786	0.026	0.825	0.184	-0.839	0.248	0.893	0.337	0.695	-0.036	0.731	0.163
	3	1.007	0.111	1.182	0.184	-0.780	0.017	0.820	0.156	-0.842	0.247	0.902	0.316	0.360	-0.109	0.757	0.165
	5	2.144	0.297	2.272	0.358	-0.845	0.022	0.915	0.204	-0.912	0.295	1.015	0.394	0.682	0.077	1.159	0.254
3	2	0.298	0.300	0.507	0.339	-0.767	0.011	0.796	0.161	-0.821	0.242	0.859	0.325	-0.090	0.092	0.377	0.181
	3	0.319	0.319	0.481	0.353	-0.797	0.016	0.842	0.166	-0.868	0.247	0.932	0.329	0.108	0.050	0.506	0.077
	5	0.514	0.370	0.636	0.418	-0.734	0.030	0.770	0.161	-0.771	0.269	0.815	0.337	0.147	0.183	0.363	0.277
4	2	-0.114	0.052	0.267	0.159	-0.658	-0.003	0.676	0.143	-0.687	0.196	0.711	0.263	-0.045	0.071	0.126	0.105
	3	-0.060	0.078	0.230	0.152	-0.677	0.023	0.694	0.128	-0.712	0.234	0.736	0.283	-0.084	0.114	0.242	0.187
	5	-0.077	0.105	0.268	0.181	-0.685	0.018	0.713	0.118	-0.721	0.228	0.761	0.270	0.116	0.090	0.200	0.142

Notes: Dynamic logit model with $\beta_1 = 0.5$, $\beta_2 = 1$, $N = 90$, and $T = 7$. Results are averaged across 50 Monte Carlo replications. See Table 1 for details.

Table 5: Classification Accuracy and CPU Time (Dynamic Model)

		NGFE					2STEPGFE							GFE				
DGP	G^0	P	R	RI	MR	CPU	P	R	RI	MR	CPU	$\widehat{G}$	Failures	P	R	RI	MR	CPU
1	2	0.50	1.0	0.50	0.46	11.06	0.51	0.91	0.51	0.90	0.49	2.33	0.82	0.53	0.55	0.54	0.38	29.60
	3	0.33	1.0	0.33	0.62	12.98	0.34	0.94	0.36	0.93	0.38	2.14	0.86	0.36	0.39	0.57	0.55	29.62
	5	0.20	1.0	0.20	0.74	16.48	0.20	0.97	0.23	0.97	0.18	2.00	0.92	0.24	0.26	0.69	0.64	29.53
2	2	0.50	1.0	0.50	0.46	8.80	0.50	0.95	0.50	0.91	0.25	2.00	0.86	0.60	0.62	0.60	0.30	21.68
	3	0.33	1.0	0.33	0.61	9.69	0.34	0.99	0.35	0.97	0.10	2.50	0.96	0.45	0.47	0.63	0.43	22.91
	5	0.20	1.0	0.20	0.74	10.05	0.23	0.97	0.28	0.92	0.37	2.33	0.82	0.36	0.46	0.74	0.54	21.09

Notes: Dynamic logit model with $\beta_1 = 0.5$, $\beta_2 = 1$, $N = 90$, and $T = 7$. Failures is the number of failures of the first step of 2STEPGFE. Results are averaged across 50 Monte Carlo replications. See Table 2 for details.

Table 6: INFERENCE FOR β_1 AND β_2 (DYNAMIC MODEL)

		NGFE						CMLE					
		SE		SD		.95		SE		SD		.95	
DGP	G^0	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
1	2	0.20	0.18	0.23	0.30	0.94	0.72	0.08	0.17	0.19	0.50	0.00	0.44
	3	0.20	0.19	0.31	0.42	0.82	0.64	0.09	0.17	0.18	0.59	0.00	0.34
	5	0.20	0.19	0.33	0.35	0.66	0.56	0.09	0.17	0.21	0.52	0.00	0.44
2	2	0.20	0.12	0.40	0.13	0.28	0.90	0.10	0.06	0.25	0.18	0.00	0.52
	3	0.23	0.13	0.62	0.15	0.30	0.72	0.12	0.07	0.25	0.16	0.00	0.60
	5	0.32	0.17	0.75	0.20	0.04	0.14	0.16	0.09	0.35	0.20	0.04	0.62
3	2	0.23	0.13	0.41	0.16	0.54	0.38	0.12	0.07	0.21	0.16	0.00	0.66
	3	0.23	0.13	0.36	0.15	0.48	0.28	0.12	0.07	0.27	0.17	0.02	0.62
	5	0.24	0.13	0.38	0.19	0.22	0.16	0.11	0.07	0.23	0.16	0.00	0.58
4	2	0.18	0.13	0.24	0.15	0.84	0.92	0.08	0.05	0.16	0.14	0.00	0.52
	3	0.18	0.13	0.22	0.13	0.88	0.92	0.08	0.05	0.15	0.13	0.00	0.68
	5	0.19	0.13	0.26	0.15	0.82	0.82	0.08	0.05	0.20	0.12	0.00	0.64

Notes: Dynamic logit model with $\beta_1 = 0.5$, $\beta_2 = 1$, $N = 90$, and $T = 7$. See Table 3 for more details.

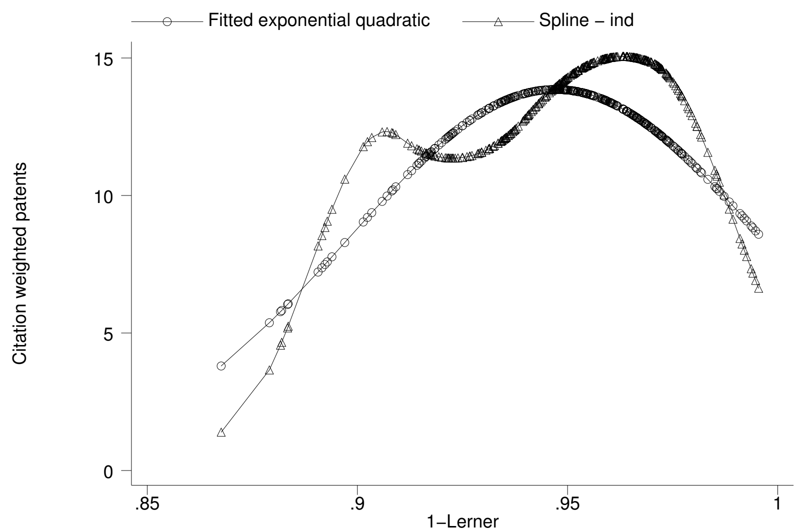
E.2 Empirical Application

Table 7: SUMMARY STATISTICS

	1-Lerner index	Citation-weighted patents	Technology gap
Mean	0.95	6.66	0.49
SD	0.02	8.43	0.16
p_{10}	0.92	0	0.28
Median	0.95	3.35	0.51
p_{90}	0.98	20.19	0.69

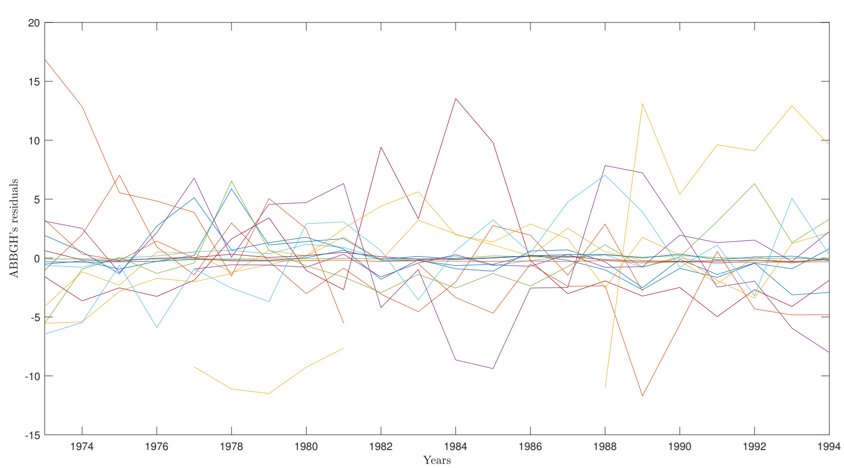
Notes: There are 17 industries and 354 observations over the time period 1973-94. See Aghion, Bloom, Blundell, Griffith, and Howitt (2005) for the exact definition of each variable.

Figure 1: Replicating ABBGH



Notes: This figure replicates Aghion, Bloom, Blundell, Griffith, and Howitt (2005)'s Figure II. Data include 17 industries of 311 firms listed on the London Stock Exchange observed between 1973 − 1994. For each industry i at year t, the prediction replaces $\widehat{\nu}_i + \widehat{\xi}_t$ with an estimated constant $\widehat{\alpha}$ (one industry and time dummies are dropped).

Figure 2: Residuals of the Two-Way Fixed Effects Poisson Model



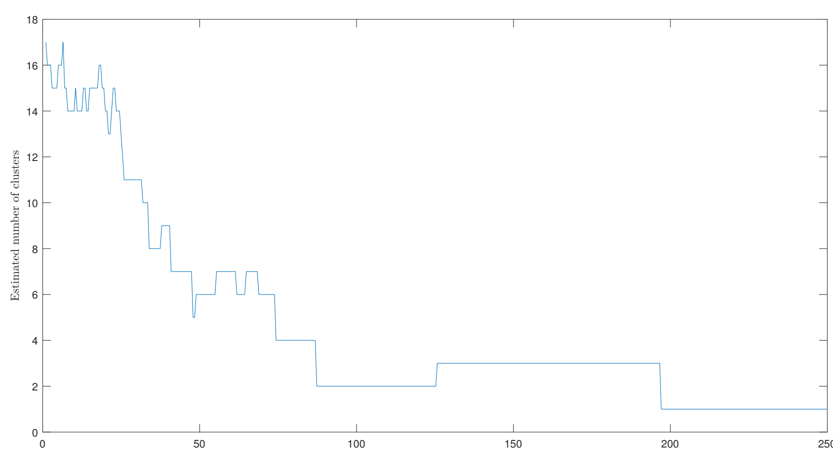
Notes: Each color represents an industry in Aghion, Bloom, Blundell, Griffith, and Howitt (2005)'s dataset. There are 17 industries observed over the period 1973-1994.

Table 8: Industries at the 2-Digit Level

SIC 2	Name
22	Metal manufacturing
23	Extraction of minerals not elsewhere specified
24	Manufacture of non-metallic mineral products
25	Chemical industry
31	Manufacture of metal goods not elsewhere specified
32	Mechanical engineering
33	Manufacture of office machinery and data processing equipment
34	Electrical and electronic engineering
35	Manufacture of motor vehicles and parts therof
36	Manufacture of other transport equipment
37	Instrument engineering
41	Food industry
42	Food, drink and tobacco manufacturing industries
43	Textile industry
47	Manufacture of paper and paper products; printing and publishing
48	Processing of rubber and plastics
49	Other manufacturing industries

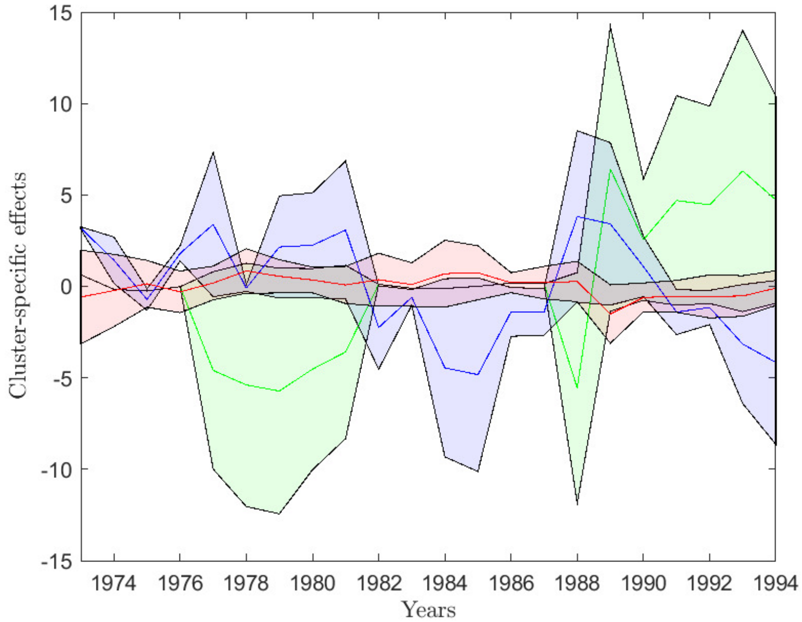
Source: 1980 Notebook of the UK Office of National Statistics available here: https://www.ons.gov.uk /methodology/classificationsandstandards/ ukstandardindustrialclassificationofeconomicactivities/uksicarchive.

Figure 3: Regularization Path of the Two-Step Pairwise Differencing Estimator



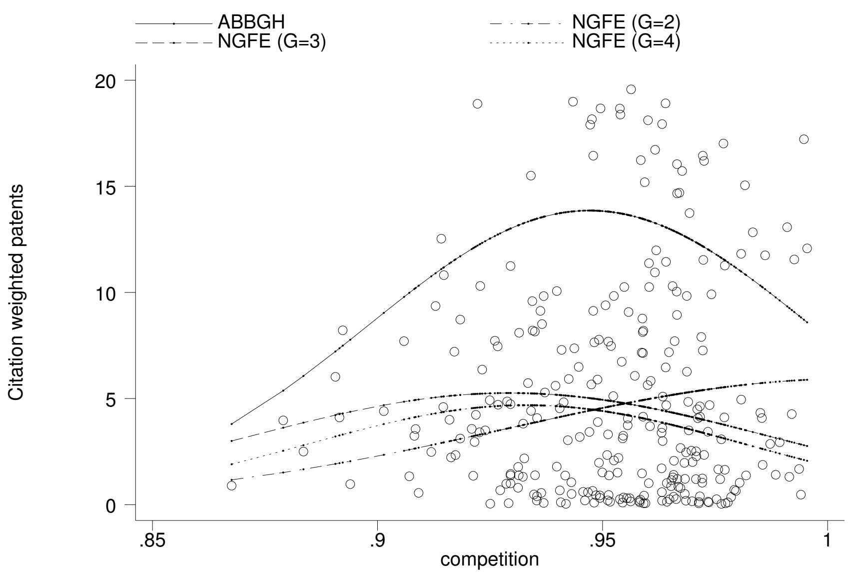
Notes: Number of estimated clusters as a function of the regularization parameter λ_2, using the pairwise distance estimator proposed in Mugnier (2022) with $\widehat{\beta}^1(\lambda_1) = 0$ (no covariates). There are 17 industries observed over the period 1973-1994.

Figure 4: Two-Step Pairwise Differencing Estimates (Three Clusters)



Notes: Each color represents an estimated cluster using the pairwise distance estimator proposed in Mugnier (2022) with $\widehat{\beta}^1(\lambda_1) = 0$ (no covariates) and $\lambda_2 \in [140, 170]$. There are 17 industries observed over the period 1973-1994.

Figure 5: Innovation and Competition Revisited: A Mildly Inverted-U Relationship



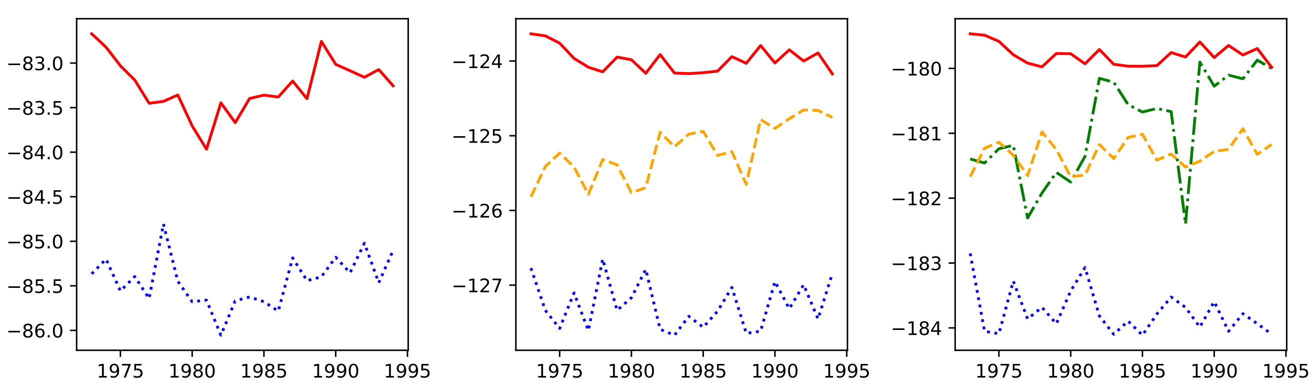
Notes: ABBGH (spe. (2) in Table 9) includes a constant and drop a time and an industry dummy (not included in the fit). NGFE (spe. (3), (4), and (5) in Table 9) does not specify a constant and averages the unobserved effects to obtain the intercept in the fit.

Table 9: The Effect of Competition on Innovation

	FE Poisson		NGFE Poisson		
Dependent variable: Citation-weighted patents$_{it}$	(1)	(2)	(3)	(4)	(5)
Competition$_{it}$	152.80***	387.46***	171.28***	273.62***	392.23***
	(55.74)	(67.74)	(71.51)	(70.21)	(70.35)
Competition squared$_{it}$	-80.99***	-204.55***	-85.15***	-147.21***	-210.19***
	(29.61)	(36.17)	(38.18)	(37.62)	(37.73)
Year effects	Yes	Yes			
Industry effects		Yes			
Time-varying clustered effects			Yes	Yes	Yes
Number of clusters			2	3	4

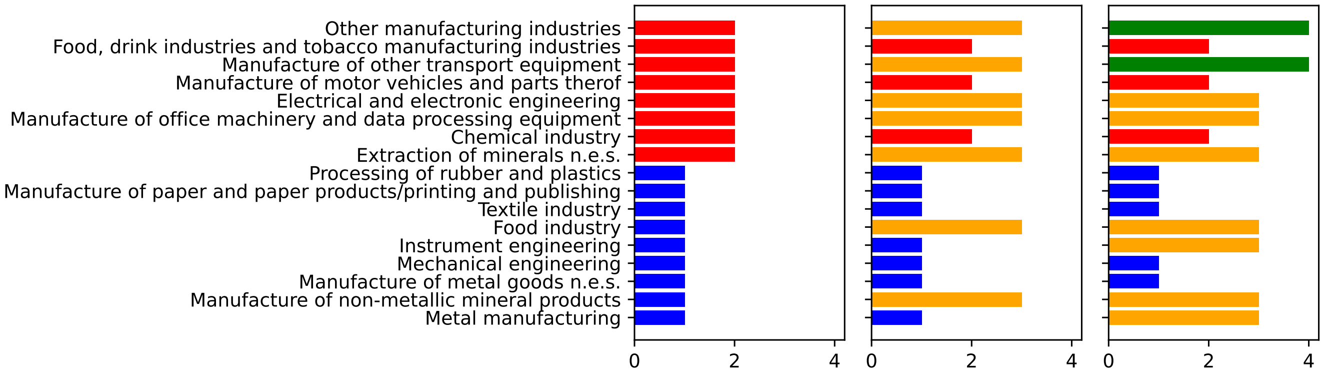
Notes: Analytical standard errors are under parentheses. The sample includes 354 observations from an unbalanced panel of 17 industries over the period 1973-1994. Competition$_{it}$ is measured by (1-Lerner index)$_{it}$ in the industry-year. NGFE estimates are computed using Lloyd's algorithm with $2,000$ random initializers. ***, **, * denote statistical significance at 1, 5, and 10% respectively.

Figure 6: Estimated Cluster-Specific Time-Varying Effects



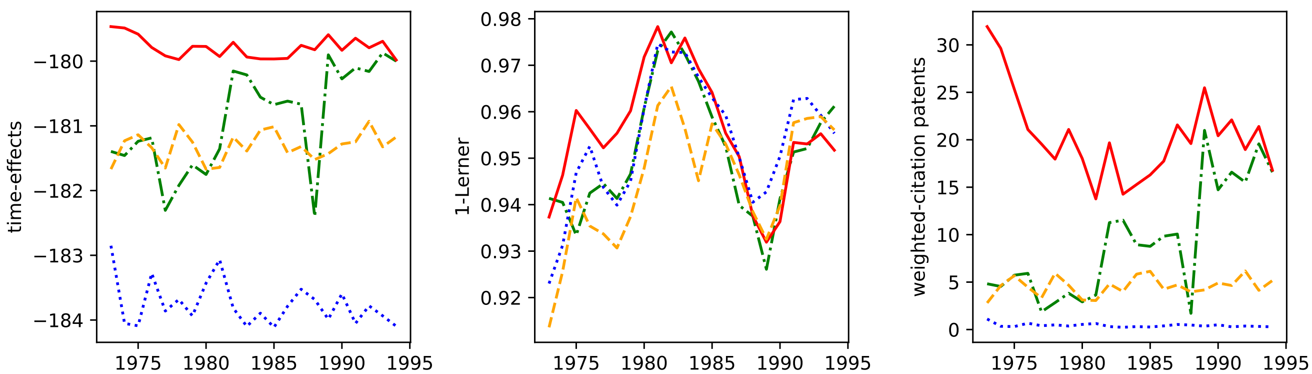
Notes: Solid red line = High-Innovation, dotted blue line = Low-Innovation, dashed orange line = Steady-Catchers, dashdotted green line = Noisy-Catchers. See Table 9 for more details.

Figure 7: Data-Driven Clusters of Industries



Notes: From left to right: $G^0 = 2, 3, 4$. Blue bar (1) = Low-Innovation, red bar (2) = High-Innovation, orange bar (3) = Steady-Catchers, green bar (4) = Noisy-Catchers.

Figure 8: Unobserved Heterogeneity, Competition, and Innovation Vary Across Time and Data-Driven Clusters



Notes: Solid red line = High-Innovation, dotted blue line = Low Innovation, dashed orange line = Steady-Catchers, dashdotted green line = Noisy-Catchers. From left to right: cluster-specific time-effects estimates ($G = 4$), results are averaged across clusters.

Table 10: The Effect of Competition on Innovation (Control Function Approach)

	FE Poisson			NGFE Poisson		
Dependent Variable: Citation-weighted patents$_{it}$	Annual	Before 1983	After 1983	Annual	Before 1983	After 1983
Competition$_{it}$	386.59***	229.18*	113.42	394.23***	265.86***	9.69
	(67.61)	(122.68)	(100.73)	(77.10)	(128.18)	(124.73)
Competition squared$_{it}$	-205.32***	-114.89*	-60.85	-212.35***	-144.18***	-9.41
	(36.11)	(66.49)	(53.37)	(41.14)	(67.95)	(67.46)
Relationship	steep inv-U	increasing		mildly inv-U	mildly inv-U	
Significance of: Competition$_{it}$, Competition squared$_{it}$	33.20	14.66	1.38			
	(0.000)	(0.001)	(0.5022)			
Significance of policy instruments	3.70	1.67	1.77	3.70	1.67	1.77
in reduced form	(0.001)	(0.192)	(0.064)	(0.001)	(0.192)	(0.064)
Significant of other instruments	5.60	3.43	2.11	5.60	3.43	2.11
in reduced form	(0.000)	(0.000)	(0.004)	(0.000)	(0.000)	(0.004)
Control functions in regression	4.38	-.61	-3.56	1.54	16.14	-2.05
	(3.51)	(6.99)	(6.13)	(2.89)	(7.05)	(3.71)
R^2 of reduced form	0.820	0.920	0.822	0.820	0.920	0.822
Year effects	Yes	Yes	Yes			
Industry effects	Yes	Yes	Yes			
Time-varying clustered effects				Yes	Yes	Yes

Notes: Competition$_{it}$ is measured by (1-Lerner index)$_{it}$ in the industry-year. The sample includes 354 observations from an unbalanced panel of 17 industries over the period 1973 to 1994 (Annual), 1973-1982 (Before 1983), or 1983-1994 (After 1983). Estimates are from a Poisson regression with industry and year fixed effects (FE) or assuming unobserved clusters of time-varying heterogeneity (NGFE) with $G = 4$ clusters of industries. Numbers in brackets are standard errors (not adjusted for the control functions). NGFE estimates are computed using Lloyd's algorithm with $2,000$ random initializers. ***, **, * denote statistical significance at 1, 5, and 10% respectively.

References

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005): "Competition and Innovation: an Inverted-U Relationship," *The Quarterly Journal of Economics*, 120(2), 701–728.

Aghion, P., J. Van Reenen, and L. Zingales (2013): "Innovation and Institutional Ownership," *American Economic Review*, 103(1), 277–304.

Aloise, D., A. Deshpande, P. Hansen, and P. Popat (2009): "NP-hardness of Euclidean sum-of-squares clustering," *Machine Learning*, 75, 245–248.

Altonji, J. G., and R. L. Matzkin (2005): "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4), 1053–1102.

Ando, T., and J. Bai (2017): "Clustering Huge Number of Financial Time Series: A Panel Data Approach With High-Dimensional Predictors and Factor Structures," *Journal of the American Statistical Association*, 112(519), 1182–1198.

Ando, T., and J. Bai (2022): "Large-scale generalized linear longitudinal data models with grouped patterns of unobserved heterogeneity," Discussion paper.

Arellano, M., and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments," in *Advances in Economics and Econometrics, Ninth World Congress.* University Press.

Arkhangelsky, D., and G. Imbens (2018): "The Role of the Propensity Score in Fixed Effect Models," .

Askenazy, P., C. Cahn, and D. Irac (2013): "Competition, R&D, and the cost of innovation: evidence for France," *Oxford Economic Papers*, 65(2), 293–311.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2021): "Matrix Completion Methods for Causal Panel Data Models," .

Athey, S., and G. W. Imbens (2006): "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 74(2), 431–497.

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71(1), 135–171.

——— (2009): "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77(4), 1229–1279.

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

Bester, C. A., and C. B. Hansen (2016): "Grouped Effects Estimators in Fixed Effects Models," *Journal of Econometrics*, 190(1), 197–208.

Boldrin, M., and D. K. Levine (2013): "The Case against Patents," *Journal of Economic Perspectives*, 27(1), 3–22.

Boneva, L., O. Linton, and M. Vogt (2015): "A semiparametric model for heterogeneous panel data with fixed effects," *Journal of Econometrics*, 188(2), 327–345.

Bonhomme, S., T. Lamadon, and E. Manresa (2022): "Discretizing Unobserved Heterogeneity," *Econometrica*, 90(2), 625–643.

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

Botosaru, I., and C. Muris (2017): "Binarization for panel models with fixed effects," CeMMAP working papers CWP31/17, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Bryant, P., and J. A. Williamson (1978): "Asymptotic Behaviour of Classification Maximum Likelihood Estimates," *Biometrika*, 65(2), 273–281.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47(1), 225–238.

——— (2010): "Binary Response Models for Panel Data: Identification and Information," *Econometrica*, 78(1), 159–168.

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20(3), S1–S13.

Chen, M., I. Fernández-Val, and M. Weidner (2021): "Nonlinear Factor Models for Network and Panel Data," *Journal of Econometrics*, 220(2), 296–324, Annals Issue: Celebrating 40 Years of Panel Data Analysis: Past, Present and Future.

Chen, M., M. Rysman, S. Wang, and K. Wozniak (2022): "An MM algorithm for Fixed Effects in Multinomial Models with an Application to Payment Choice in Grocery Stores," .

Cheng, X., F. Schorfheide, and P. Shao (2021): "Clustering for Multi-Dimensional Heterogeneity," Discussion paper, University of Pennsylvania.

Chetverikov, D., and E. Manresa (2021): "Spectral and Post-Spectral Estimators for Grouped Panel Data Models," Discussion paper.

Correa, J. A. (2012): "Innovation and competition: An unstable relationship," *Journal of Applied Econometrics*, 27(1), 160–166.

Davezies, L., X. D'Haultfoeuille, and M. Mugnier (2020): "Fixed Effects Binary Choice Models with Three or More Periods," .

de Chaisemartin, C., and X. D'Haultfœuille (2020): "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110(9), 2964–96.

de Paula, A. (2020): "Econometric Models of Network Formation," .

Deb, P., and P. Trivedi (1997): "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics*, 12, 313–36.

Dhaene, G., and K. Jochmans (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82(3), 991–1030.

D'Haultfoeuille, X. (2010): "A new instrumental method for dealing with endogenous selection," *Journal of Econometrics*, 154(1), 1–15.

D'Haultfoeuille, X., S. Hoderlein, and Y. Sasaki (2021): "Testing and relaxing the exclusion restriction in the control function approach," *Journal of Econometrics*.

Dzemski, A., and R. Okui (2018): "Confidence set for group membership," .

Evdokimov, K. (2011): "Nonparametric Identification of a Nonlinear Panel Model with Application to Duration Analysis with Multiple Spells," .

Fernández-Val, I., and M. Weidner (2016): "Individual and Time Effects in Nonlinear Panel Models With Large N, T," *Journal of Econometrics*, 192(1), 291–312.

——— (2018): "Fixed Effects Estimation of Large-T Panel Data Models," *Annual Review of Economics*, 10(1), 109–138.

Forgy, E. (1965): "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, 21, 768–780.

Gao, J., K. Xia, and H. Zhu (2020): "Heterogeneous Panel Data Models With Cross-sectional Dependence," *Journal of Econometrics*, 219(2), 329–353, Annals Issue: Econometric Estimation and Testing: Essays in Honour of Maxwell King.

Gao, W. Y., M. Li, and S. Xu (2022): "Logical differencing in dyadic network formation models with nontransferable utilities," *Journal of Econometrics*.

Gilbert, R. (2006): "Looking for Mr. Schumpeter: Where Are We in the Competition–Innovation Debate?," *Innovation Policy and the Economy*, 6, 159–215.

Giraud, C. (2014): *Introduction to High-Dimensional Statistics*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Goodfellow, I. J., Y. Bengio, and A. Courville (2016): *Deep Learning*. MIT Press, Cambridge, MA, USA.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 52(3), 701–720.

Graham, B. S. (2017): "An Econometric Model of Network Formation With Degree Heterogeneity," *Econometrica*, 85(4), 1033–1063.

Griffith, R., and J. Van Reenen (2021): "Product market competition, creative destruction and innovation," LSE Research Online Documents on Economics 113816, London School of Economics and Political Science, LSE Library.

Gu, J., and S. Volgushev (2019): "Panel Data Quantile Regression With Grouped Fixed Effects," *Journal of Econometrics*, 213(1), 68 – 91, Annals: In Honor of Roger Koenker.

Hahn, J., and H. R. Moon (2010): "Panel Data Models With Finite Number of Multiple Equilibria," *Econometric Theory*, 26(3), 863–881.

Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hansen, B. E. (2008): "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24(3), 726–748.

Hashmi, A. (2013): "Competition and Innovation: The Inverted-U Relationship Revisited," *The Review of Economics and Statistics*, 95(5), 1653–1668.

Hausman, J., B. H. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R & D Relationship," *Econometrica*, 52(4), 909–938.

Heckman, J. (1981): "Heterogeneity and State Dependence," in *Studies in Labor Markets*, pp. 91–140. National Bureau of Economic Research, Inc.

Heckman, J., and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2), 271–320.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983): "Stochastic blockmodels: First steps," *Social Networks*, 5(2), 109–137.

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

Hsiao, C. (2015): *Analysis of Panel Data*. Cambridge University Press.

Ibragimov, I. A. (1956): "On the Composition of Unimodal Distributions," *Theory of Probability & Its Applications*, 1(2), 255–260.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58(1), 71–120.

Janys, L., and B. Siflinger (2021): "Mental Health and Abortions among Young Women: Time-varying Unobserved Heterogeneity, Health Behaviors, and Risky Decisions," .

Ke, Y., J. Li, and W. Zhang (2016): "STRUCTURE IDENTIFICATION IN PANEL DATA ANALYSIS," *The Annals of Statistics*, 44(3), 1193–1233.

Klein, R. W., and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), 387–421.

Kock, A. B. (2016): "Oracle Inequalities for High Dimensional Panel Data Models," .

Kock, A. B., and H. Tang (2019): "UNIFORM INFERENCE IN HIGH-DIMENSIONAL DYNAMIC PANEL DATA MODELS WITH APPROXIMATELY SPARSE FIXED EFFECTS," *Econometric Theory*, 35(2), 295–359.

Krantz, S., and H. Parks (2002): *A Primer of Real Analytic Functions*, Advanced Texts Series. Birkhäuser Boston.

Lesigne, E., and D. Volný (2001): "Large deviations for martingales," *Stochastic Processes and their Applications*, 96(1), 143–159.

Lewbel, A. (2014): "38An Overview of the Special Regressor Method," in *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press.

Liu, R., Z. Shang, Y. Zhang, and Q. Zhou (2020): "Identification and estimation in panel models with overspecified number of groups," *Journal of Econometrics*, 215(2), 574–590.

Lloyd, S. P. (1982): "Least squares quantization in PCM," *IEEE transactions on information theory*, 28(2), 129–137.

MacQueen, J. B. (1967): "Some Methods for Classification and Analysis of MultiVariate Observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, ed. by L. M. L. Cam, and J. Neyman, vol. 1, pp. 281–297. University of California Press.

McLachlan, G. J., and D. Peel (2000): *Finite mixture models*, vol. 299 of *Probability and Statistics – Applied Probability and Statistics Section*. Wiley, New York.

Merlevède, F., M. Peligrad, and E. Rio (2011): "A Bernstein type inequality and moderate deviations for weakly dependent sequences," *Probability Theory and Related Fields*, 151(3), 435–474.

Moon, H., and M. Weidner (2017): "Dynamic Linear Panel Regression Models With Interactive Fixed Effects," *Econometric Theory*, 33(1), 158–195.

Moon, H. R., and M. Weidner (2015): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," *Econometrica*, 83(4), 1543–1579.

——— (2019): "Nuclear Norm Regularized Estimation of Panel Regression Models," .

Mugnier, M. (2022): "A Simple and Computationally Trivial Estimator for Grouped Fixed Effects Models," .

Mugnier, M., and A. Wang (2022): "Identification and (Fast) Estimation of Large Nonlinear Panel Models with Two-Way Fixed Effects," Discussion paper.

Neyman, J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16(1), 1–32.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

Pratt, J. W. (1981): "Concavity of the Log Likelihood," *Journal of the American Statistical Association*, 76(373), 103–106.

Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Denmarks Paedagogiske Institute.

Rio, E. (2000): *Théorie asymptotique des processus aléatoires faiblement dépendants*. Springer, Berlin, – Heidelberg, – New York.

Roth, J., and Rambachan (2022): "A More Credible Approach to Parallel Trends," Discussion paper.

Saggio, R. (2012): "Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models," Master's thesis, Center for Monetary and Financial Studies.

Su, L., Z. Shi, and P. C. B. Phillips (2016): "Identifying Latent Structures in Panel Data," *Econometrica*, 84(6), 2215–2264.

Su, L., X. Wang, and S. Jin (2019): "Sieve Estimation of Time-Varying Panel Data Models With Latent Structures," *Journal of Business & Economic Statistics*, 37(2), 334–349.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

Vershynin, R. (2019): "High-Dimensional Probability," .

Vogt, M., and O. Linton (2017): "Classification of non-parametric regression functions in longitudinal data models," *Journal of the Royal Statistical Society Series B*, 79(1), 5–27.

von Luxburg, U. (2007): "A Tutorial on Spectral Clustering," .

Vytlacil, E., and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75(3), 757–779.

Wang, W., and L. Su (2021): "Identifying Latent Group Structures in Nonlinear Panels," *Journal of Econometrics*, 220(2), 272–295, Annals Issue: Celebrating 40 Years of Panel Data Analysis: Past, Present and Future.

Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data, second edition*, Econometric Analysis of Cross Section and Panel Data. MIT Press.

Yu, L., J. Gu, and S. Volgushev (2022): "Group structure estimation for panel data – a general approach," .

Zeleneev, A. (2020): "Identification and Estimation of Network Models with Nonparametric Unobserved Heterogeneity," Discussion paper.

Zhang, Y., H. J. Wang, and Z. Zhu (2019): "Quantile-regression-based clustering for panel data," *Journal of Econometrics*, 213(1), 54–67, Annals: In Honor of Roger Koenker.