Essays in Panel Data Econometrics

Martin Mugnier

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAG008

Thèse de doctorat

ENSAE

IP PARIS

Essays in Panel Data Econometrics

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Sciences économiques

Thèse présentée et soutenue à Palaiseau, le 19 juin 2023, par

MARTIN MUGNIER

Composition du Jury :

Anna Simoni Professeur, CREST-ENSAE-École Polytechnique (UMR 9194)	Présidente
Stéphane Bonhomme Professeur, University of Chicago	Rapporteur
Koen Jochmans Professeur, Toulouse School of Economics	Rapporteur
Frank Windmeijer Professeur, University of Oxford	Examinateur
Xavier D'Haultfœuille Professeur, CREST-ENSAE (UMR 9194)	Directeur de thèse

Le hasard est une divinité – plus compliquée, plus exigeante, plus secrète qu'aucune.

Julien Gracq, *Un beau ténébreux*

Dans une seule idée d'un créateur vivent mille nuits d'amour oubliées.

Rainer Maria Rilke, *Lettres à un jeune poète*

Acknowledgements

In *La Formation de l'esprit scientifique*, Gaston Bachelard once wrote

> Avant tout, il faut savoir poser des problèmes. Et quoi qu'on dise, dans la vie scientifique, les problèmes ne se posent pas d'eux-mêmes. C'est précisément ce *sens du problème* qui donne la marque du véritable esprit scientifique. Pour un esprit scientifique, toute connaissance est une réponse à une question. S'il n'y a pas eu de question, il ne peut y avoir connaissance scientifique. Rien ne va de soi. Rien n'est donné. Tout est construit.

Writing the last words of this dissertation, I now better understand what Bachelard meant. Asking the right questions is fine art, answering them is lifelong learning.

To have put me on track, encouraged me to cultivate scientific rigor, transmitted to me his taste for try and fail, his sense of ethics and modesty, to have pushed me to develop my intellectual curiosity, keep courage, and sharpen my critical thinking, in a few words, to have taught me how to do important, sustainable and enjoyable high-quality research, I want to thank my Ph.D. supervisor Xavier D'Haultfœuille. Working with and learning from him has been a chance for me, and he remains a solid source of inspiration for my early career as an academic. I hope our collaboration will continue in the future.

Other people have greatly inspired me to pursue excellence and develop a strong taste for curiosity and rigorous thinking. Early in my studies, I think of Christelle Fairise and Pascal Vanhove. Later on, Daniel-Li Chen, Matthieu Lequien, Antonin Bergeaud, Loriane Py, Jérémy L'Hour, Christophe Gaillac, and Stéphane Bonhomme. I am indebted to Anna Simoni who, maybe unconsciously, triggered my will to pursue a Ph.D. when I was a young graduate student attending her Econometrics course at the École Polytechnique. There I discovered, not without some apprehension but also with excitement, I still had a long way to go beyond my nascent understanding of econometrics. Together with Victor-Emmanuel Brunel, I thank her for being part of my annual internal thesis committee.

I thank Pierre Alquier, Christophe Giraud, and the many great professors from the Master in Statistics and Machine Learning in Orsay who trusted me and supported my Ph.D. project at the intersection of econometrics, statistics, and machine learning. My research benefits a lot from what I have learned from them. I thank Éric Gautier for his helpful advice at early stages.

Research is a collective activity. I am grateful to my exceptional co-authors: Xavier D'Haultfœuille, Laurent Davezies, Ao Wang, and Jérémy L'Hour. I have learned a lot from them, and I am confident that we will have opportunities to work together again in the future. I thank and wish a long life to the CREST-PSE econometric group, in which I intend to continue to actively participate.

I want to thank the examiners, referees, and members of my dissertation jury who kindly accepted to evaluate this work. It is an honor to have you as a jury.

I thank the French Ministry of Higher Education, Research and Innovation and the Agence Nationale de la Recherche for funding most of this research. Although somewhat rippled by elitism, the French Grandes Écoles system, I believe, fosters excellence and provides the necessary facilities and environment to produce international standard research. A pure product of this system, I am grateful to it, in particular the École Normale Supérieure Paris-Saclay (my alma mater) and ENSAE.

Beyond my beloved students and daily colleagues at CREST, fellow doctoral students and faculty, I want to thank the directors of doctoral studies, Thibaud Vergé and Alessandro Riboni, as well as members of the administration and technical support, particularly, Arnak Dalalyan, Eliane Madelaine, Édith Verger, Murielle Jules, Tristan Duchenne, Djamila Gherarbi, Lyza Racoon, Fanda Traoré, Leyla Marzuk, Weronika Leduc, Philippe Pinczon du Sel, and Teddy Arrif. CREST benefits a lot from such an efficient and joyful staff.

I am grateful to Stéphane Bonhomme for his warm welcome at the University of Chicago, his continuing guidance and support. Like Xavier, I rarely met someone so enthusiastic about research, friendly, and curious. I thank the faculty members of UChicago for encouraging comments, in particular, Guillaume Pouliot, Alex Torgovitsky, Max Tabord-Meehan, Azeem Shaikh, and Jim Heckman. Thomas B., Art Institute of Chicago, Smartbar, Lou, Cécile, Jonas, Eyo, Jiarui, Thomas W., Hugo, Sasha, Léon, Kathleen, and Émilien: my stay at the Kenneth C. Griffin Department of Economics would probably not have been so great without having met you.

Suzette Tanis-Plant and Ted Eames provided valuable help in the process of writing official documents in the language of Shakespeare. I am very grateful to them.

Special thanks go to my family and friends, especially the team of Ramonvillois. These years have been special to me and required a lot of abnegation. Being surrounded by you was precious. I particularly thank Clément G. for our multiple chats on Gaussian processes, L-statistics, A. Grothendieck, J. Gracq, P. Modiano, uncertainty quantification, and sensitivity analysis, for our lunches on the "platal" and our (too rare) mojitos at every corners of the 13th and 14th arrondissements.

I owe so much to my parents. Somehow, they are the first responsible for who I am today. I thank them for all their love and the nurturing environment which have allowed me to emancipate intellectually.

When my brother and I were young, we were reluctant to go to those endless *art et essai* movies, strange art exhibitions and plays, or long walks in the countryside and lost villages on Sunday afternoons. We didn't understand why we didn't have television at home. Today I understand the value of all this. Nino, the thought of our music jams – sometimes together with unfinished proofs – made me wake up everyday and continue. Anne, you made me read the best books of my life. Eric, you kind of taught me mathematics in the end.

Research is love. I am grateful to all the people who shared their love with me.

Contents

List of Figures

List of Tables

To Eric, Anne, and Nino. To all the musicians, poets, and writers who have brighten the Day 'N' Nite

Résumé Substantiel en Français

Cette thèse propose de nouvelles méthodes économétriques pour l'analyse des données de panel. Les données de panel incluent les données longitudinales, où les unités statistiques telles que des individus, firmes ou pays sont observées à plusieurs dates dans le temps, ainsi que les données hiérarchiques, par exemple, les caractéristiques d'élèves affectés à des professeurs, d'employés travaillant pour certaines firmes, de biens et services vendus sur plusieurs marchés, ou de flux commerciaux entre pays au cours du temps (panel en trois dimensions).

Un problème important en économie appliquée demeure la prise en compte de l'hétérogénéité inobservée (Heckman, 1981). Les unités peuvent différer systématiquement selon des caractéristiques inobservées par l'économètre, ce qui rend difficile l'identification de paramètres d'intérêts. Par exemple, il est plausible que les individus aux capacités de concentration et d'organisation élevées choisissent en moyenne de poursuivre des études plus longues et obtiennent des salaires élevés qu'ils auraient pu obtenir indépendamment de leur niveau d'éducation : les différences moyennes de salaire observées par niveaux de diplômes ne rendent pas bien compte des rendements moyens de l'éducation (voir Abowd et al., 1999). Ce "biais de variable omise" menace la plupart des études exploitant des données non-expérimentales, l'expérience "idéale" telle qu'une assignation aléatoire d'individus au sein d'un groupe de contrôle et de test étant bien souvent impossible en économie (e.g., Angrist and Pischke, 2008).

À l'instar de la notion de parcimonie en statistiques en grande dimension (*sparsity*), les données de panel offrent de nombreuses opportunités à l'économètre pour prendre en compte une hétérogénéité inobservée de grande dimension à condition que celle-ci ait une structure simple. Cette tension entre flexibilité (grande dimension des modèles) et parcimonie de la structure est au cœur de la modélisation des données de panels et de la formulation des modèles économétriques modernes. À titre d'illustration, supposons que la capacité individuelle de l'individu i puisse être résumée par une variable scalaire non-observée $A_i \in \mathbb{R}$ et que le salaire à l'instant t, noté $W_{it} \in \mathbb{R}$, soit une fonction linéaire de la capacité A_i, du nombre d'années d'études achevées à l'instant t, noté $E_{it} \in \mathbb{N}$, et d'un terme d'erreur spécifique à l'individu et à la période $U_{it} \in \mathbb{R}$:

$$W_{it} = A_i + \beta E_{it} + U_{it}, \quad \mathbb{E}(U_{it}|A_i, E_{i1}, \ldots, E_{iT}) = 0, \quad i = 1, \ldots, N,\ t = 1, \ldots, T.$$

Le paramètre $\beta \in \mathbb{R}$ représente l'effet marginal moyen d'une année d'études supplémentaire (de $e-1$ à e) sur le salaire d'un individu de niveau de capacité a :

$$\beta = \mathbb{E}(W_{it}|A_i = a, E_{it} = e) - \mathbb{E}(W_{it}|A_i = a, E_{it} = e-1).$$

Si $\mathbb{E}(A_i|E_{it}) \neq 0$, il se peut que les individus à haute capacité gagnent des salaires plus élevés et poursuivent une éducation plus longue. Sous des hypothèses de régularité standards, le coefficient obtenu en utilisant une régression linéaire des moindres carrés ordinaires regroupée des salaires sur les années d'études est un estimateur biaisé de l'effet marginal moyen β. En raison de la linéarité du modèle et de la constance temporelle de la capacité, la variable de capacité non-observée A_i peut être facilement éliminée par une différenciation temporelle de premier ordre du modèle :

$$\Delta W_{it} = \beta \Delta E_{it} + \Delta U_{it}, \quad \mathbb{E}(\Delta U_{it}|\Delta E_{i2}, \ldots, \Delta E_{iT}) = 0, \quad i = 1, \ldots, N,\ t = 2, \ldots, T,$$

avec $\Delta W_{it} = W_{it} - W_{it-1}$, $\Delta E_{it} = E_{it} - E_{it-1}$ et $\Delta U_{it} = U_{it} - U_{it-1}$. Sous des conditions standard, une régression linéaire des moindres carrés ordinaires regroupée des différences temporelles de salaires sur les différences temporelles d'années d'études fournit un estimateur sans biais et convergent de β dès que le nombre d'individus N diverge (e.g., Mundlak, 1961; Wooldridge, 2010).

La simplicité de cette procédure ne doit pas cacher sa puissance. Le résultat d'identification pour β est fort car le modèle semi-paramétrique impose très peu d'hypothèses sur l'hétérogénéité non-observée : la distribution conditionnelle de la capacité non-observée A_i étant donné le nombre d'années d'éducation complétées E_{it} est entièrement non-restreinte (au-delà du modèle conditionnel pour $W_{it}|E_{i1}, \ldots, E_{iT}, A_i$), la distribution du terme d'erreur U_{it} n'est pas spécifiée, et la corrélation temporelle des termes d'erreur $(U_{it})_{t=1,\ldots,T}$ n'est pas un problème de premier ordre.

Les choses deviennent radicalement différentes dans les modèles non-linéaires avec une hétérogénéité non-observée variante dans le temps. La non-linéarité peut résulter de comportements de maximisation d'utilité et la capacité peut varier de manière endogène au cours de la vie d'un travailleur. Premièrement, la plupart des objets d'intérêt ne sont que partiellement identifiés (e.g., Chamberlain, 2010; Chernozhukov et al., 2013; Davezies et al., 2022). Deuxièmement, l'estimation et l'inférence deviennent plus difficiles (Arellano and Hahn, 2007).

Une question centrale de cette thèse est de déterminer si des approches de différenciation computationnellement simples s'appliquent aux modèles plus complexes, tels que des modèles non-linéaires, dynamiques et avec hétérogénéité inobservée variant dans le temps. Cette ligne de recherche complète les idées de différenciation fonctionnelle et les généralisations développées notamment dans Athey and Imbens (2006), Bonhomme (2012) et Hoderlein and White (2012). En particulier, cette thèse propose de nouvelles méthodes pour les données de panel visant à identifier, estimer et effectuer une inférence statistique sur des paramètres causaux dans certains modèles économétriques (non-linéaires) avec hétérogénéité inobservée. Les nouvelles méthodes

permettent des relations fonctionnelles non-linéaires entre les variables endogènes et exogènes, des interactions entre les paramètres non-observés spécifiques à l'individu et spécifiques au temps, des distributions d'erreurs flexibles et des routines de calcul rapides. Les panels courts et larges sont étudiés.

Le premier chapitre, écrit en collaboration avec Xavier D'Haultfœuille et Laurent Davezies, généralise les résultats de Johnson (2004) et Chamberlain (2010) en montrant que le paramètre de pente dans un modèle de choix binaire statique avec trois périodes ou plus peut être identifié de manière ponctuelle même si les chocs idiosyncratiques ne suivent pas une distribution logistique. Nous fournissons une restriction de moment conditionnelle, qui peut être utilisée pour obtenir un estimateur asymptotiquement normal au taux paramétrique, lorsque le nombre d'unités diverge vers l'infini, en appliquant la méthode des moments généralisés (GMM). Nous illustrons cette nouvelle méthode en revisitant la relation entre les déficits budgétaires et les réélections étudiée dans Brender and Drazen (2008). L'effet significatif et positif du déficit budgétaire sur la probabilité de réélection est robuste à une relaxation de l'hypothèse logistique.

Le deuxième chapitre, rédigé en collaboration avec Ao Wang, présente de nouveaux résultats d'identification et des conditions suffisantes pour une classe de modèles non-linéaires à doubles effets fixes avec des coefficients hétérogènes pour les panels larges et longs. Nous proposons une procédure d'estimation rapide basée sur un algorithme de descente de gradient coordonnée par coordonnée de type Gauss-Siedel qui exploite la séparabilité additive des effets fixes. Dans le cas semi-paramétrique, nous démontrons l'équivalence numérique de notre méthode avec l'estimateur du maximum de vraisemblance, reportons des gains de calcul importants sans perte de précision au regard des routines existantes (par exemple, logitfe/probitfe dans Stata) et revisitons deux applications empiriques en innovation (Aghion et al., 2013) et commerce international (Helpman et al., 2008). Nous trouvons une hétérogénéité significative des coefficients de pente relatifs aux variables indépendantes dans chacun des modèles.

Les troisième et quatrième chapitres traitent d'un cas particulier de modèles factoriels, dans lesquels les facteurs individuels sont supposés discrets. Cette hypothèse génère une structure de groupe qui peut rationaliser une grande variété de configurations économiques (par exemple, clubs de pays, partenaires commerciaux, types de consommateurs, de biens et d'actifs financiers). Le troisième chapitre propose un nouvel estimateur en deux étapes pour le modèle linéaire, qui présente plusieurs avantages théoriques et computationnels. En résolvant un programme d'optimisation convexe et en utilisant une procédure de regroupement agglomérative, nous généralisons Bonhomme and Manresa (2015) et montrons que le paramètre de pente commun, les effets fixes et le nombre de groupes peuvent être estimés de manière convergente sans borne supérieure connue sur le nombre de groupes, tout en réduisant la complexité algorithmique à l'ordre du cube du nombre d'unités contre une complexité exponentielle pour l'estimateur reposant sur l'algorithme des k-means. Le quatrième chapitre étend certains de ces résultats à une classe de modèles non-linéaires à variable dépendante

discrète.

Le cinquième et dernier chapitre, écrit en collaboration avec Xavier D'Haultfœuille et Jérémy L'Hour, démontre la normalité asymptotique d'estimateurs définis comme moyennes empiriques de la transformation d'une fonction de répartition empirique par un processus quantile empirique, sous des hypothèses bien plus faibles que celles connues actuellement. Un exemple populaire est l'estimateur "Changes-in-Changes" proposé dans Athey and Imbens (2006). Nous obtenons de nouveaux résultats en utilisant la théorie des L-statistiques (Shorack and Wellner, 1986). Certains de nos résultats intermédiaires peuvent avoir un intérêt indépendant. Des simulations de Monte Carlo suggèrent que ces nouvelles hypothèses sont nécessaires.

Introduction

> *– Ah oui ! Les économistes ! dit Kostanjoglo, sans l'écouter et avec une expression sarcastique... De fameux imbéciles qui en mènent d'autres et ne voient pas plus loin que leur nez ! Des ânes qui montent en chaire et mettent des lunettes... Tas d'idiots !*
>
> Nicolas Gogol, *Les Âmes mortes*

Despite Kostanjoglo's radical view, economists have long drawn inferences by combining economic theory with data science (Haavelmo, 1943, 1944). To understand complex relationships between economic variables, to predict counterfactual events, or to evaluate public policies, they build structural models, run randomized controlled experiments, and carefully exploit observational data. Most rely on several branches of applied mathematics to answer scholarly, business, and policy-relevant questions.

A common concern in applied work is unobserved heterogeneity. Economic agents, unlike gaz particles, make decisions based on characteristics unobserved to the researcher (Heckman, 1981). Unobserved heterogeneity may render identification of key parameters of interest difficult. High-ability individuals plausibly stay longer in school and would earn high wages anyway: observed differences in wages across degrees do not identify marginal returns to education (e.g., Abowd et al., 1999). This so-called omitted variable bias is a spectre haunting applied economics, where "ideal" randomized experiments are rarely feasible (e.g., Angrist and Pischke, 2008).

Panel data provide opportunities to account for high-dimensional unobserved heterogeneity. Panel data is longitudinal data, where units of observations such as individuals, firms, or countries are observed at different points in time; or hierarchical data, for instance, goods purchased on different markets, students assigned to teachers, or trade flows between countries over time (a three dimensional panel). Akin to the notion of sparsity in high-dimensional statistics (e.g., Giraud, 2014), high-dimensional ability may have a low-dimensional underlying structure relative to the number of observations. As a short illustration, suppose that individual i's ability can be summarized by an unobserved scalar variable $A_i \in \mathbb{R}$, while her wage at time t, denoted by $W_{it} \in \mathbb{R}$, is a linear function of ability A_i, her completed years of education at time t, denoted by $E_{it} \in \mathbb{N}$, and some individual-time specific error term $U_{it} \in \mathbb{R}$:

$$W_{it} = A_i + \beta E_{it} + U_{it}, \quad \mathbb{E}(U_{it}|A_i, E_{i1}, \dots, E_{iT}) = 0, \quad i = 1, \dots, N,\ t = 1, \dots, T.$$

The parameter $\beta \in \mathbb{R}$ is the average marginal effect of one additional year of education (from $e-1$ to e) on wages given ability level a:

$$\beta = \mathbb{E}(W_{it}|A_i = a, E_{it} = e) - \mathbb{E}(W_{it}|A_i = a, E_{it} = e-1).$$

If $\mathbb{E}(A_i|E_{it}) \neq 0$, high-ability individuals might have higher wages and longer education, and – under standard regularity conditions – the coefficient obtained by running a pooled Ordinary Least Squares (OLS) linear regression of wages on years of educations is a biased estimate of the average marginal effect β. Because of linearity and time-constant ability, the unobserved ability variable A_i is eliminated by a first-order time-differencing transformation of the model:

$$\Delta W_{it} = \beta \Delta E_{it} + \Delta U_{it}, \quad \mathbb{E}(\Delta U_{it}|\Delta E_{i2}, \ldots, \Delta E_{iT}) = 0, \quad i = 1, \ldots, N,\ t = 2, \ldots, T,$$

where $\Delta W_{it} = W_{it} - W_{it-1}$, $\Delta E_{it} = E_{it} - E_{it-1}$, and $\Delta U_{it} = U_{it} - U_{it-1}$. Under standard conditions, a pooled OLS linear regression of time differences in wages on time differences in years of education is unbiased and consistent for β as long as the number of individuals N diverges (e.g., Mundlak, 1961; Wooldridge, 2010).

The simplicity of the manipulation should not undermine its usefulness. The identification result for β is strong and desirable as the semi-parametric model imposes very few assumptions on the unobserved heterogeneity: the conditional distribution of unobserved ability A_i given completed years of education E_{it} is fully unrestricted (beyond the conditional model for $W_{it}|E_{i1}, \ldots, E_{iT}, A_i$), the distribution of the error term U_{it} is not specified, and serial correlation in $(U_{it})_{t=1,\ldots,T}$ is not a first-order issue.

Things become radically different in nonlinear models with time-varying unobserved heterogeneity. Nonlinearities may arise from utility-maximization behaviors, and ability may vary endogeneously during a worker's life. First, most objects of interest are only partially identified (e.g., Chamberlain, 2010; Chernozhukov et al., 2013; Davezies et al., 2022). Second, estimation and inference become more challenging (Arellano and Hahn, 2007). A central question to this dissertation is whether computationally straightforward differencing approaches apply to more complicated models, e.g., nonlinear, dynamic, and time-varying unobserved heterogeneity models. This line of research complements the functional differencing ideas and generalizations developed in, e.g., Athey and Imbens (2006), Bonhomme (2012), and Hoderlein and White (2012).

This dissertation proposes new panel data methods to identify, estimate, and perform statistical inference on causal parameters in (nonlinear) econometric models with unobserved heterogeneity. The new methods allow for nonlinear functional relationships between endogenous and exogenous variables, interactions between individual-specific and time-specific unobserved parameters, flexible error terms distributions, and fast computation routines. Both short and large panels are studied.

The first chapter, written jointly with Xavier D'Haultfœuille and Laurent Davezies, generalizes Johnson (2004) and Chamberlain (2010)'s results by showing that

the slope parameter in a static binary choice model with three periods or more can be point identified even if idiosyncratic shocks do not follow the restrictive logistic distribution. We provide a conditional moment restriction, which can be used to obtain an asymptotically normal estimator at the parametric rate, when the number of units diverges to infinity, by applying the Generalized Method of Moments (GMM). We illustrate this new method by revisiting the relationship between budget deficits and reelections studied in Brender and Drazen (2008). The significant positive effect of budget deficits on the probability of reelection is robust to departures from the logistic assumption.

The second chapter, written jointly with Ao Wang, presents new identification results and sufficient conditions for a class of nonlinear two-way fixed effects models with heterogeneous coefficients for large and long panels. We provide a fast estimation procedure based on a Gauss-Siedel coordinate-wise gradient descent algorithm which exploits additive separability in the fixed effects. In the semiparametric case, we prove the numerical equivalence of our method to the maximum likelihood estimator, we report considerable gains in execution time without loss in precision with respect to existing packages (e.g., logitfe/probitfe in Stata), and we revisit two empirical applications in innovation (Aghion et al., 2013) and international trade (Helpman et al., 2008). We find significant heterogeneity in estimated slopes for the independent variables in each case.

The third and fourth chapters consider a special case of factor models, in which individual factor loadings are assumed discrete. This assumption generates a group structure that can rationalize a wide variety of economic settings (e.g., clubs of countries, trading partners, types of consumers, goods, financial assets). The third chapter proposes a new two-step estimator for the linear model, which has several theoretical and computational advantages. By solving a convex optimization program and using an agglomerative clustering procedure, we generalize Bonhomme and Manresa (2015) and show that the common slope parameter, the fixed effects, and the number of groups can be consistently estimated without a known upper bound on the number of groups while reducing algorithmic complexity to the order of the cube of the number of units against an exponential complexity for the estimator relying on the k-means algorithm. The fourth chapter extends some of these results to a class of nonlinear models for discrete outcomes.

The fifth and last chapter, written jointly with Xavier D'Haultfœuille and Jérémy L'Hour, proves the asymptotic normality of estimators defined as empirical means of the transform of an empirical cumulative distribution functions by an empirical quantile process under much weaker assumptions than what is currently known. One popular example is the "Changes-in-Changes" estimator proposed in Athey and Imbens (2006). We obtain these new results by using results from the theory of L-statistics (see, e.g. Shorack and Wellner, 1986). Some of our intermediate results may have independent interest. Monte Carlo simulations suggest that our assumptions cannot be improved.

This thesis mostly contributes to econometric theory. This subfield of economics consists not only in developing a better mathematical understanding of existing economic models and estimation procedures, but also in proposing new methodologies and statistical guarantees to quantify uncertainty in one's conclusions. An economic model is a set of restrictions on the phenomenon under study. Once economic theory provides such restrictions, formal conditions can be established and statistical procedures be developed to take decisions, learn, and test further economic theories using experimental or observational data (e.g., Gourieroux and Monfort, 1995; Manski, 2021). While some investigations require brand new economic theories, economic researchers otherwise often combine popular existing models and statistical tools. A caricatural example is OLS. OLS lie at the heart of instrumental variables regression, Local Average Treatment Effects (LATE) analysis (Angrist and Imbens, 1995; Imbens and Angrist, 1994), estimation of trade gravity equations (Helpman et al., 2008), longitudinal studies of wage returns to education (Abowd et al., 1999), etc. As many other procedures, OLS can be analyzed in the very general framework of models defined by moment and inequality restrictions (e.g., Hansen, 1982). Hence, it seems reasonable to conduct methodological research independently from any practical economic questions.[1] This direction has seemingly been taken by an important body of the theoretical econometric literature, and we follow it here. Also, this illustrates that few problems lie at the heart of many economic questions: selection, unobserved heterogeneity, and functional form restrictions.

To conclude this introduction, a few remarks are in order. To which extent can a theory explain human facts? Are there quantitative laws that economists should disclose? Are models still useful in the Big Data era? This dissertation does not address these difficult epistemological questions. Nevertheless, the latter help understand the roots of the problems considered here. The economics we are interested in, *the social science concerned chiefly with the way society chooses to employ its resources, which have alternative uses, to produce goods and services for present and future consumption*, rarely focuses on "all human beings" producing "all types of goods and services", but rather addresses well-delimited questions involving agents plausibly very close to mechanically applying optimization routines, whether or not it is profit or welfare maximization (e.g., firms, financial institutions, governments, NGOs); large aggregates (e.g., macroeconomic consumption, investment, production, trade flows between countries); or human beings acting in highly competitive markets or situations for which financial considerations are of primary order (e.g., private investors, highly-qualified job seekers, tax payers). There, the existence of natural laws holding "on average" that economists should disclose seems hardly controversial. Outside of this scope, e.g., when the focus lies on individual-level decisions *per se* which involve a lot of considerations that plausibly go beyond purely optimizing behaviors, let say at the household level for which some methods developed in this

[1] That theoretical econometrics is not merely "mathematical statistics" or "data science" simply follows from the fact that "econ" is still there: questions, assumptions, and theorems generally differ between these neighboring fields.

dissertation could find important applications (e.g., models of choices), economics is often criticized for its imperialism. This is perhaps not without reason.

Relying on utility-maximization principles as a basis for economic analysis, however, is certainly not the illusory dream of a few ultra-liberal economists convinced that human beings can be reduced to cold calculating machines, nor is it the desesperate attempt of a few lost physicists convinced that there is not much difference between modeling human consumption behaviors and missile trajectories using optimal control theory. A model is an idealization, which allows to reach an agreement regarding the validity of a proposed solution. When so-called Nobel Prize Gary Becker's *A Theory of Marriage* (Becker, 1973, 1974) modeled the decision to marry as the result of a utility-maximization process, eternally single philosophy master Arthur Schopenhauer had already confessed, in *L'Art de se connaître soi-même*, "*Ne plus pouvoir disposer librement de ma propre personne est un mal bien plus grand que l'avantage qui naîtrait potentiellement du gain d'une autre*". Cost-benefit analysis is therefore infinitely more delicate than what a rough picture may suggest.

An economic theory formulates stable assumptions on agents behaviors and possible states of nature with the purpose of deriving informative propositions. Abstracting from an agent's budget, if utility derived from car consumption at period $t \in \{0, 1, ...\}$ is $u_{c,t} \in \mathbb{R}$, utility from bike consumption is $u_{b,t} \in \mathbb{R}$, utility from the outside-option of commuting by feet is $u_{f,t} \in \mathbb{R}$, and an agent chooses at each time period what is best for her in terms of derived utility, then a resulting optimal action plan $(a_t^*)_{t\in\mathbb{N}} \in \{c, b, f\}^{\mathbb{N}}$ is such that

$$u_{a_t^*,t} \geq u_{a,t}, \qquad \text{for all } (a, t) \in \{c, b, f\} \times \mathbb{N}.$$

The above equation simply summarizes that the agent chooses what she prefers to do at each period. How does the agent derive utility? How to deal with multiple heterogeneous agents that differ based on unobserved dimensions absent from the data? These are important questions underlying many everyday economic problems. Unfortunately, the abstract cary-all summary measure of "utility" is rarely fully observed as a function of both observed economic treatments of interest x_t (e.g., distance to workplace, income, type of job, socio-economic class of parents, etc) and unobserved individual-specific attributes α_t (e.g., deep preference for an ecological mode of transportation) that may systematically be related to x_t. Since many important economic questions tantamount to learn from the data functionals of population marginal derivatives $\partial u_{a,t}(x_t, \alpha_t)/\partial x_t$ or counterfactuals $u_{a,t}(x_t', \alpha_t')$ in order to quantify (causal) consequences of changes in the economic environement, and because the perfect randomized controlled experiment is rarely feasible in economics (e.g., because of costs and ethical considerations), it is important to develop rich models for observational data which parcimoniously but flexibly account for multidimensional sources of unobserved heterogeneity.

Panel data allow the researcher to take into account unobserved heterogeneity of a

particular form when estimating a causal model. For instance, under the assumption that motivation, ability, and familial background remain stable over time and any other factors correlated with participation and future labor market outcome evolve similarly for different individuals over time, some mathematically well-defined average causal effect are identified from the data and can be estimated at parametric rates (see, e.g., de Chaisemartin and D'Haultfœuille, 2020). Such common trend restrictions are pervasive, but preclude many types of nonlinearities between the variables of interest.

In nonlinear panel data models, identification and estimation of key (causal) parameters in the semi-parametric case remain challenging because of the incidental parameters problem Neyman and Scott (1948). In this thesis, we aim at contributing to the literature on fixed-T, large-N asymptotics and large-T, large-N asymptotics by considering several special cases where the unobserved heterogeneity has some low-dimensional structure (it lies on some low-dimensional manifold) so that high-dimensional statistical tools such as for instance penalization or clustering may provide satisfying solutions. Applications range from microeconometrics to macroeconometrics. Reference textbooks for panel data (micro)econometrics are Hsiao (2014); Pesaran (2015); Wooldridge (2010).

Chapter 1

Fixed Effects Binary Choice Models with Three or More Periods

La pensée commence par un retrait du monde, une absence, c'est-à-dire une absence à une présence, à un « présent » toujours pressant et urgent puisqu'il doit se définir avec rigueur par les besoins de la vie et par les exigences du corps. (...) Ainsi, penser, c'est d'une manière ou d'une autre, par un choix plus ou moins libre, exercer une capacité plus ou moins indépendante, se détacher.

Frédéric Worms, *Penser à quelqu'un*

Abstract: We consider fixed effects binary choice models with a fixed number of periods T and regressors without a large support. If the time-varying unobserved terms are i.i.d. with known distribution F, Chamberlain (2010) shows that the common slope parameter is point identified if and only if F is logistic. However, he only considers in his proof $T = 2$. We show that the result does not generalize to $T \geq 3$: the common slope parameter can be identified when F belongs to a family including the logit distribution. Identification is based on a conditional moment restriction. Under restrictions on the covariates, these moment conditions lead to point identification of relative effects. If $T = 3$ and mild conditions hold, GMM estimators based on these conditional moment restrictions reach the semiparametric efficiency bound. Finally, we illustrate our method by revisiting Brender and Drazen (2008).[1]

[1]This chapter is based on a co-authored paper with Laurent Davezies (CREST-ENSAE) and Xavier D'Haultfœuille (CREST-ENSAE). It has been accepted for publication in *Quantitative Economics* in 2022.

1.1 Introduction

In this chapter, we revisit the classical binary choice model with fixed effects. Specifically, let T denote the number of periods and let us suppose to observe, for individual i, $(Y_{it}, X_{it})_{t=1,\dots,T}$ with

$$Y_{it} = \mathbb{1}\{X_{it}'\beta_0 + \gamma_i - \varepsilon_{it} \geq 0\} \tag{1.1.1}$$

where $\beta_0 \in \mathbb{R}^K$ is unknown and $\varepsilon_{it} \in \mathbb{R}$ is an idiosyncratic shock. The nonlinear nature of the model and the absence of restriction on the distribution of γ_i conditional on $X_i := (X_{i1}', \dots, X_{iT}')'$ renders the identification of β_0 difficult. Rasch (1960) shows that if the $(\varepsilon_{it})_{t=1,\dots,T}$ are i.i.d. with a logistic distribution, a conditional maximum likelihood can be used to identify and estimate β_0. Chamberlain (2010) establishes a striking converse of Rasch's result: if the $(\varepsilon_{it})_{t=1,\dots,T}$ are i.i.d. with distribution F and the support of X_i is bounded, β_0 is point identified only if F is logistic. Other papers have circumvented such an impossibility result by either considering large support regressors (see in particular Honore and Lewbel, 2002; Manski, 1987) or allowing for dependence between the shocks (see Magnac, 2004).

It turns out, however, that Chamberlain (2010) only proves his result for $T = 2$. And in fact, we show that his result does not generalize to $T \geq 3$. Specifically, we consider distributions F satisfying

$$\frac{F(u)}{1-F(u)} = \sum_{k=1}^{\tau} w_k \exp(\lambda_k u) \quad \text{or} \quad \frac{1-F(u)}{F(u)} = \sum_{k=1}^{\tau} w_k \exp(-\lambda_k u), \tag{1.1.2}$$

with $T \geq \tau + 1$, $(w_1, \dots, w_\tau) \in (0, \infty)^\tau$, and $1 = \lambda_1 < \dots < \lambda_\tau$. We study the identification of β_0, assuming that $\lambda := (\lambda_1, \dots, \lambda_\tau)$ is known. The weights $w_1, \dots, w_\tau$ remain unknown, thus allowing for much more flexibility on the distribution of ε_{it} than in the logit case. In particular, it may either be left- or right-skewed, platykurtic or leptokurtic. Our main insight is that for any F satisfying (1.1.2), a conditional moment restriction holds. We also obtain some results on the corresponding identified set B. For instance, if, roughly speaking, X_i is continuous, we show that B includes at most $T! - 1$ points (2 if $T = 3$) and relative marginal effects are point identified. Note that Johnson (2004) considers the same family with $\tau = 2$ and $T = 3$. However, he does not study the general case and does not show any formal identification result based on the corresponding moment conditions.

Obviously, the conditional moment condition can be used to construct GMM estimators. This means, in particular, that $\sqrt{n}$-consistent estimation is possible beyond the logit case when $T > 2$, overturning again the impossibility results of Chamberlain (2010) and Magnac (2004). Further, we show that if $T = 3$ and mild additional restrictions hold, the optimal GMM estimator based on our conditional moment conditions reaches the semiparametric efficiency bound of the model. Hence, at least when $T = 3$, these moment conditions contain all the information of the model.

Finally, we showcase the empirical relevance of our approach by studying whether budget deficits and economic growth affect reelection, revisiting Brender and Drazen (2008). The authors investigate this issue using simple and fixed effects logit models. However, the assumption of logistic errors is not warranted, so we consider whether the results are robust to this assumption on the unobserved terms. Our results suggest that the relative effects of budget deficits and economic growth or other variables are fairly robust to the logistic assumption.

This chapter is related to the seminal work of Bonhomme (2012), who develops a unified approach for models where the conditional distribution of $(Y_1, \dots, Y_T)$ given (X_i, γ_i) is parametrized by β_0, but no restriction on the distribution of $\gamma_i|X_i$ is imposed. In such set-ups, he shows that the identification and estimation of β_0 depends on the existence of functions $m \neq 0$ satisfying

$$\mathbb{E}[m(Y, X, \beta_0)|X, \gamma] = 0.$$

This approach has been fruitfully applied to the dynamic logit model by Kitazawa (2022) and Honoré and Weidner (2020). Our work may be seen as yet another application of this approach, focusing on static models but dropping the logistic assumption.

The remainder of the chapter is organized as follows. Section 1.2 describes the moment condition we use for identification of β_0 and establishes some properties of the identified set based on these moments. Section 1.3 discusses GMM estimation of β_0, links it with the semiparametric efficiency bound of the model and discusses the case of unbalanced panel data. Section 1.4 is devoted to the application. Section 1.5 concludes. All the proofs are collected in the appendix.

1.2 Identification

1.2.1 The Model and Moment Conditions

We drop the subscript i in the absence of ambiguity and let $Y = (Y_1', \dots, Y_T')'$, $X = (X_1', \dots, X_T')'$, $X_t = (X_{1,t}, \dots, X_{K,t})'$, $X_{k,\cdot} = (X_{k,1}, \dots, X_{k,T})'$, $X_{-k} = (X_{k',t})_{k' \neq k, t=1,\dots,T}$, $X_{k,-t} = (X_{k,s})_{s \neq t}$, and $X_{-k,t} = (X_{k',t})_{k' \neq k}$. $\text{Supp}(X) \subset \mathbb{R}^{KT}$ denotes the support of the random variable X. For any set $A \subset \mathbb{R}^p$ (for any $p \geq 1$), we let $A^* := A \backslash \{0\}$ and denote by $|A|$ the cardinal of A. Hereafter, we maintain the following conditions.

Assumption 1.2.1 (Binary choice panel model) *Equation (1.1.1) holds and:*

1. *(X, γ) and $(\varepsilon_t)_{1 \leq t \leq T}$ are independent and the $(\varepsilon_t)_{1 \leq t \leq T}$ are i.i.d. with a known cumulative distribution function (cdf) F.*
2. *For all (k, t), $\mathbb{E}[X_{k,t}^2] < \infty$.*
3. *$\beta_0 \in \mathbb{R}^{K*}$.*

The first condition is also considered in Chamberlain (2010). The second condition is a standard moment restriction on the covariates. Finally, we exclude in the third condition the case $\beta_0 = 0$ here. This case can be treated separately, as the following proposition shows.

Proposition 1.2.1 *Suppose that Assumption 1.2.1 holds, F is strictly increasing on $\mathbb{R}$ and there exist $(t, t') \in \{1, \dots, T\}^2$ such that $\mathbb{E}[(X_t - X_{t'})(X_t - X_{t'})']$ is nonsingular. Then $\beta_0 = 0$ if and only if*

$$\mathbb{P}(Y_t = 1, Y_{t'} = 0 | Y_t + Y_{t'} = 1, X_t, X_{t'}) = \frac{1}{2} \quad a.s. \tag{1.2.1}$$

Condition (1.2.1) can be tested by a specification test on the nonparametric regression of $D = Y_t(1 - Y_{t'})$ on $(X_t, X_{t'})$, conditional on the event $Y_t + Y_{t'} = 1$. See, e.g., Bierens (1990) or Hong and White (1995).

Turning to identification on $\mathbb{R}^{K*}$, we first recall the impossibility result of Chamberlain (2010). We say below that F is logistic if $G(u) := F(u)/(1 - F(u)) = w \exp(\lambda u)$ for some $(w, \lambda) \in \mathbb{R}^{+*2}$.

Theorem 1.2.2 *Suppose that $T = 2$, X_t includes $\mathbb{1}\{t = 2\}$, Assumption 1.2.1.1 holds, F is strictly increasing on $\mathbb{R}$ with bounded, continuous derivative and Supp(X) is compact. If F is not logistic, there exists $\beta_0 \in \mathbb{R}^{K*}$, a distribution of $\gamma|X$ and an open ball $B \subset \mathbb{R}^K$ such that β_0 is not identified compared to $\beta \in B$.*

This result implies in particular that when $T = 2$ and F is not logistic, relative effects β_{0j}/β_{0k}, for k such that $\beta_{0k} \neq 0$, may not be identified. Such relative effects are important as they are equal to relative marginal effects if both $X_{j,t}$ and $X_{k,t}$ are continuous. If only $X_{k,t}$ is continuous (say), $-\beta_{0j}/\beta_{0k}$ still corresponds to a compensating variation.[2]

The key step in Chamberlain's proof is that if β_0 is identified for all data generating process satisfying the restrictions of the theorem, the conditional probabilities (conditional on X and γ) of the four possible trajectories for (Y_1, Y_2) are necessarily affinely dependent. Moreover, by letting $|\gamma|$ tend to infinity, the stable trajectories $(0, 0)$ and $(1, 1)$ disappear from this relationship. This leads to the following functional equation for G:

$$\psi_1(\alpha)G(u) + \psi_2(\alpha)G(u + \alpha) = 0, \tag{1.2.2}$$

[2]To see the first point, note that under Assumptions 1.2.1-1.2.2,

$$\mu_{k,t}(x) := \frac{\partial \mathbb{P}(Y_t = 1 | X_{k,t} = x_{k,t}, X_{k,-t} = x_{k,-t}, X_{-k} = x_{-k})}{\partial x_{k,t}} = \beta_{0k} \mathbb{E}[F'(x_t'\beta_0 + \gamma) | X = x]$$

and thus $\mu_{j,t}(x)/\mu_{k,t}(x) = \beta_{0j}/\beta_{0k}$. Also, $-\beta_{0j}/\beta_{0k}$ corresponds to the change in $X_{k,t}$ necessary to keep $\mathbb{P}(Y_t = 1 | X_t, \alpha)$ constant when $X_{j,t}$ increases by one unit.

for all $u \in \mathbb{R}$, α in an open subset of $\mathbb{R}$ and some functions $\psi_1(\cdot), \psi_2(\cdot)$ such that for all α, $(\psi_1(\alpha), \psi_2(\alpha)) \neq (0,0)$. The result follows by noting that the solutions necessarily have the form $u \mapsto w \exp(\lambda u)$.

Equation (1.2.2) relies on the time dummy variable $\mathbb{1}\{t = 2\}$. However, the proof of Theorem 2 of Chamberlain (2010) shows that even without such a dummy variable, (1.2.2) is necessary for the semiparametric efficiency bound not to be zero, or, equivalently, for the existence of regular, root-n consistent estimators of β_0. In this case, α corresponds to $(x_2 - x_1)'\beta_0$, for (x_1, x_2) in a set of positive measure.

In any case, the same reasoning with $T = 3$ leads to the following equation for G:

$$\begin{aligned}\psi_1(\boldsymbol{\alpha})G(u) + \psi_2(\boldsymbol{\alpha})G(u+\alpha_1) + \psi_3(\boldsymbol{\alpha})G(u+\alpha_2) + \psi_4(\boldsymbol{\alpha})G(u)G(u+\alpha_1) \\ +\psi_5(\boldsymbol{\alpha})G(u)G(u+\alpha_2) + \psi_6(\boldsymbol{\alpha})G(u+\alpha_1)G(u+\alpha_2) = 0,\end{aligned} \tag{1.2.3}$$

for all $u \in \mathbb{R}$, $\boldsymbol{\alpha} := (\alpha_1, \alpha_2)$ in an open subset of $\mathbb{R}^2$ and some functions $\psi_k(\cdot)$, $k = 1, \ldots, 6$, such that for for all $\boldsymbol{\alpha}$, $(\psi_1(\boldsymbol{\alpha}), \ldots, \psi_6(\boldsymbol{\alpha})) \neq (0, \ldots, 0)$. We now have $6 = 2^3 - 2$ terms instead of just $2 = 2^2 - 2$, and thus we can expect to have other solutions than just $u \mapsto w \exp(\lambda u)$. And indeed, one can check that if G has the form $u \mapsto w_1 \exp(\lambda_1 u) + w_2 \exp(\lambda_2 u)$, we can construct $(\psi_1(\boldsymbol{\alpha}), \psi_2(\boldsymbol{\alpha}), \psi_3(\boldsymbol{\alpha})) \neq (0,0,0)$ such that (1.2.3) holds, with $\psi_4(\boldsymbol{\alpha}) = \psi_5(\boldsymbol{\alpha}) = \psi_6(\boldsymbol{\alpha}) = 0$. Similarly, if $1/G$ has the form $u \mapsto w_1 \exp(\lambda_1 u) + w_2 \exp(\lambda_2 u)$, we can construct $(\psi_4(\boldsymbol{\alpha}), \psi_5(\boldsymbol{\alpha}), \psi_6(\boldsymbol{\alpha})) \neq (0,0,0)$ such that (1.2.3) holds, with $\psi_1(\boldsymbol{\alpha}) = \psi_2(\boldsymbol{\alpha}) = \psi_3(\boldsymbol{\alpha}) = 0$. Note that there may still be other solutions to (1.2.3) that are increasing and have a limit of ∞ (resp. 0) at ∞ (resp. at $-\infty$). The question of identifying all such solutions is left for future research.

Generalizing this reasoning to any $T > 2$, we see that combinations of at most $T - 1$ exponential functions satisfy the functional restrictions tantamount to (1.2.3) and which render identification of β_0 possible. This suggests that identification may be achieved for the corresponding family of distribution, which we now formally introduce. Hereafter, Λ_τ denotes a subset of $\{(\lambda_1, \ldots, \lambda_\tau) \in \mathbb{R}^\tau : 1 = \lambda_1 < \ldots < \lambda_\tau\}$.

Assumption 1.2.2 ("Generalized" logistic distributions)[3] *There exist known* $\tau \in \{1, \ldots, T-1\}$ *and* $\lambda := (\lambda_1, \ldots, \lambda_\tau)' \in \Lambda_\tau$ *and unknown* $w := (w_1, \ldots, w_\tau)' \in (0, \infty)^\tau$ *such that:*

$$\begin{aligned} \textit{Either} \quad & F(u)/(1-F(u)) = \sum_{j=1}^{\tau} w_j \exp(\lambda_j u) \quad \textit{(First type)}, \\ \textit{or} \quad & (1-F(u))/F(u) = \sum_{j=1}^{\tau} w_j \exp(-\lambda_j u) \quad \textit{(Second type)}. \end{aligned}$$

Fixing $\min\{\lambda_1, \ldots, \lambda_\tau\}$ to 1 is without loss of generality, as we can always multiply β_0, γ_i and ε_{it} by this factor. If F is of the second type, then one can show that the cdf

[3] Though we use the same name, our family of distributions should not be confused with those introduced by Balakrishnan and Leung (1988) and Stukel (1988).

of $-\varepsilon_{it}$ is of the first type. Thus, up to changing (Y_t, X_t) into $(1 - Y_t, -X_t)$, we can assume without loss of generality, as we do afterwards, that F is of the first type. We shall see that $\tau + 1$ periods are sufficient to achieve identification. Hence, we assume, again without loss of generality, that $T = \tau + 1$: if $T > \tau + 1$, we can always focus on $\tau + 1$ periods.

Before describing our identification strategy of β_0 when F is a generalized logistic distribution, two remarks are in order. First, we obtain our results below irrespective of the vector w.[4] Hence, in contradistinction with the fixed effect logistic model, we do not fix the distribution of ε, but simply impose that it belongs to a family of distributions indexed by two parameters. Members of this family differ in particular by their skewness and kurtosis. In linear regressions, the residuals are often found to have a skewed distribution with either positive or negative excess kurtosis. Then, there is no reason why the latent variables corresponding to Y_{it} would not exhibit a similar pattern. On the other hand, we do fix λ. Identification of λ could also be of interest but is not addressed in this work.

Now, the idea behind the identification of β_0 is to construct a function $m \neq 0$ such that $\mathbb{E}[m(Y, X, \beta_0)|X, \gamma] = 0$ almost surely. Thus, as mentioned in the introduction, we apply Bonhomme (2012)'s general idea of functional differencing. The function m is related to the functions ψ_k in (1.2.3) when $T = 3$, and the generalization of (1.2.3) when $T > 3$. For any $x = (x_1', \ldots, x_T')' \in \mathbb{R}^{KT}$, let $x_s^{-t} = x_s$ if $s < t$, $x_s^{-t} = x_{s+1}$ else. We let

$$M_t(x; \beta) = (-1)^{t+1} \det \begin{pmatrix} \exp(\lambda_1 x_1^{-t\prime}\beta) & \ldots & \exp(\lambda_1 x_{T-1}^{-t\prime}\beta) \\ \vdots & & \vdots \\ \exp(\lambda_{T-1} x_1^{-t\prime}\beta) & \ldots & \exp(\lambda_{T-1} x_{T-1}^{-t\prime}\beta) \end{pmatrix}.$$

Then define, for any $(y, x, \beta) \in \{0, 1\}^T \times \text{Supp}(X) \times \mathbb{R}^{K*}$,

$$m(y, x; \beta) := \sum_{t=1}^{T} \mathbb{1}\{y_t = 1, y_{t'} = 0 \ \forall t' \neq t\} M_t(x; \beta).$$

Our first result shows that m, indeed, satisfies a conditional moment restriction:

Theorem 1.2.3 *If Assumptions 1.2.1-1.2.2 hold, we have, almost surely,*

$$\mathbb{E}[m(Y, X; \beta_0)|X, \gamma] = \mathbb{E}[m(Y, X; \beta_0)|X] = 0. \tag{1.2.4}$$

Theorem 1.2.3 shows there exists a known moment condition which potentially identifies β_0 in a more general model than the logistic one. Also, as the number of periods T increases, the class of distributions F for which β_0 can be point identified increases. This is consistent with the idea that if $T = \infty$, β_0 is point identified

[4] We do impose however that all the components of w are non-zero, for normalization purposes. Otherwise, the model with $w = (w_1, 0)$ and β_0, for instance, would be equivalent to the model with $w = (0, w_1)$ and β_0/λ_2. A similar issue arises with, e.g., $w = (w_1, w_2, 0)$ if $\lambda_3/\lambda_2 = \lambda_2/\lambda_1$.

for any F, by using variations in X_t of a single individual. Note however that the class of generalized logistic distribution is not dense for the set of all cdf's: any cdf F belonging to the closure of this class should be such that either $F/(1-F)$ or $(1-F)/F$ is convex. In Section 1.7, we exhibit an even more general class of distributions (when weights w's are known) that does not suffer this limitation and for which we show that moment conditions exist for T sufficiently large. Theorem 1.2.3 also complements the results of Chernozhukov et al. (2013) showing that bounds on β_0 for general F shrink quickly as T increases.

Theorem 1.2.3 holds with $T = \tau + 1 = 2$. In such a case, the conditional moment condition can be written

$$\mathbb{E}\left[\mathbb{1}\{Y_1 > Y_2\}\exp(X_2'\beta_0) - \mathbb{1}\{Y_2 > Y_1\}\exp(X_1'\beta_0)|X\right] = 0.$$

This conditional moment generates the first-order conditions of the maximization of the theoretical conditional likelihood, since these the first-order conditions are equivalent to

$$\mathbb{E}\left[\frac{(X_1 - X_2)}{\exp(X_1'\beta_0) + \exp(X_2'\beta_0)}\left(\mathbb{1}\{Y_1 > Y_2\}\exp(X_2'\beta_0) - \mathbb{1}\{Y_2 > Y_1\}\exp(X_1'\beta_0)\right)\right] = 0.$$

1.2.2 Necessary and Sufficient Conditions for Identification

The discussion above implies that with $T = \tau + 1 = 2$, β_0 is identified by (1.2.4) as soon as $\mathbb{E}\left[(X_1 - X_2)(X_1 - X_2)'\right]$ is nonsingular. We now turn to the more difficult case where $T - 1 = \tau > 1$. Let B denote the identified set of β_0 obtained with our conditional moment conditions, namely

$$B := \left\{b \in \mathbb{R}^{K*} : \mathbb{E}[m(Y, X; b)|X] = 0 \text{ a.s.}\right\}.$$

We also denote by $B_k := \{b_k : \exists b = (b_1, \ldots, b_k, \ldots, b_K) \in B\}$ $(k = 1, \ldots, K)$ the identified set of β_{0k}. Our first result shows that B is included in a set depending on the distribution of X only. To define this set, let us introduce

$$D_j(x; b) := \det\begin{pmatrix} \exp(\lambda_j x_1'\beta_0) & \ldots & \exp(\lambda_j x_T'\beta_0) \\ \exp(\lambda_1 x_1' b) & \ldots & \exp(\lambda_1 x_T' b) \\ \vdots & & \\ \exp(\lambda_{T-1} x_1' b) & \ldots & \exp(\lambda_{T-1} x_T' b) \end{pmatrix}$$

and, for all $b \in \mathbb{R}^{K*}$, let

$$\mathcal{D}(b) = \Big\{x \in \text{Supp}(X) : \max_{j=1,\ldots,T-1} D_j(x; b) > \min_{j=1,\ldots,T-1} D_j(x; b) \geq 0 \\ \text{or } \min_{j=1,\ldots,T-1} D_j(x; b) < \max_{j=1,\ldots,T-1} D_j(x; b) \leq 0\Big\}.$$

Because $D_j(x;\beta_0) = 0$ for all $x \in \text{Supp}(X)$, we have $\mathbb{P}(X \in \mathcal{D}(\beta_0)) = 0$. The following lemma shows that B is actually included in the set of b's satisfying this property.

Lemma 1.2.4 *Suppose that Assumptions 1.2.1-1.2.2 hold. Then,*

$$B \subset \widetilde{B} := \left\{ b \in \mathbb{R}^{K*} : \mathbb{P}(X \in \mathcal{D}(b)) = 0 \right\}.$$

This result follows because the moment condition can be written as a weighted sum of the $D_j(x;b)$'s, with positive weights. It shows that β_0 is identified if for all nonzero $b \neq \beta_0$, we can find some $x \in \text{Supp}(X)$ such that all nonzero $D_j(x;b)$ have the same sign, and the set of such nonzero determinants is not empty.

The set $\widetilde{B}$ is convenient in that it does not depend on the unknown distribution of $\gamma|X$; but it is hard to characterize in general. Nevertheless, we are able to obtain results under either of the conditions below.

Assumption 1.2.3 *For all $k \in \{1,\dots,K\}$, $\mathbb{P}\left(|\{X_{k,1},\dots,X_{k,T}\}| = T, X_{-k} = 0\right) > 0$.*[5]

Assumption 1.2.4 *There exists $(s,t,x) \in \{1,\dots,T\}^2 \times \mathbb{R}^K$, $s < t$ and a neighborhood V of x such that Supp$(X) \cap \left[\mathbb{R}^{(s-1)K} \times V \times \mathbb{R}^{(t-s-1)K} \times V \times \mathbb{R}^{(T-t)K}\right]$ has a non-empty interior.*

The first assumption corresponds to a case where all components of X are discrete. It imposes that for all k and t, the support of $X_{k,t}$ includes 0 and at least $T-1$ additional elements. Because we can always replace $X_{k,\cdot}$ by $X_{k,\cdot} - c_k$ for any $c_k \in \mathbb{R}^T$, the condition $0 \in \text{Supp}(X_{k,t})$ for all k,t holds as long as $\cap_{t=1}^T \text{Supp}(X_{k,t})$ is not empty (for all k). The second condition imposes that all components of X_t are continuous. It also imposes that for at least two periods s and t, $\text{Supp}(X_s) \cap \text{Supp}(X_t)$ is not empty. This last condition holds for instance if $(X_t)_{t\geq 1}$ is strictly stationary.

Theorem 1.2.5 *Suppose that Assumptions 1.2.1-1.2.2 hold. Then:*

1. *If Assumption 1.2.3 also holds, $|B| < \infty$ and $B_k \subset \{c\beta_{0k} : c \in \{0\} \cup (1/\lambda_{T-1}, \lambda_{T-1})\}$.*

2. *If Assumption 1.2.4 also holds,*

$$B \subset \widetilde{B} \subset R := \{c\beta_0 : c \in (1/\lambda_{T-1}, \lambda_{T-1})\}. \tag{1.2.5}$$

Moreover, $|B| \leq T! - 1$ and $|B| \leq 2$ when $T = 3$. All relative effects β_{0j}/β_{0k}, for k such that $\beta_{0k} \neq 0$, are point identified.[6]

[5] When $K = 1$, the condition $X_{-k} = 0$ should simply be omitted.

[6] The set of indices k such that $\beta_{0k} \neq 0$ is identified since by (1.2.5), $B_k = \{0\}$ when $\beta_{0k} = 0$, and $0 \notin B_k$ otherwise.

Whether Assumption 1.2.3 or 1.2.4 holds, Theorem 1.2.5 shows that under-identification is at most finite, namely $|B| < \infty$. This implies that β_0 is locally identified in the sense that there exists a neighborhood of β_0 in which the unique solution to the equation $\mathbb{E}[m(Y, X; b)|X] = 0$ is $b = \beta_0$. Further, the first result of Theorem 1.2.5 shows that with discrete regressors satisfying Assumption 1.2.3, the "length" of the identified set on β_{0k}, defined as

$$\max_{(b_{1k}, b_{2k}) \in B_k^2} |b_{1k} - b_{2k}|,$$

cannot exceed $\beta_{0k}(\lambda_{T-1} - 1/\lambda_{T-1})$ if $0 \notin B_k$. Note that under Assumption 1.2.3, we can actually identify whether or not $\beta_{0k} = 0$ without relying on our conditional moments, since the sign of β_{0k} is equal to that of $\mathbb{E}[Y_t - Y_s | X_{-k,s} = X_{-k,t}, X_{k,t} > X_{k,s}]$. The second result on continuous regressors is stronger. It shows that if Assumption 1.2.4 holds, β_0 is identified up to a scale c, with c belonging at most to $(1/\lambda_{T-1}, \lambda_{T-1})$. This directly implies point identification of relative marginal effects. The second result also states that B includes at most $T! - 1$ points, and even only 2 points when $T = 3$. Importantly, all these result hold for any possible distribution of $\gamma|X$. Thus, point identification may actually hold for many distributions of $\gamma|X$, a point we shall come back to below.

The proof of Theorem 1.2.5 relies on the following ideas. In the first case, when $b_k \notin \{c\beta_{0k} : c \in \{0\} \cup (1/\lambda_{T-1}, \lambda_{T-1})\}$, we construct a subset of $\text{Supp}(X)$ of positive probability such that all nonzero $D_j(x; b)$ have the same sign. The result then follows by Lemma 1.2.4. We use a similar reasoning to prove (1.2.5). To establish the upper bounds on $|B|$, we exploit the fact that the family of exponential functions $(v \mapsto \exp(\zeta_k v))_{k=1,\ldots,K}$ with distinct coefficients ζ_k forms a Chebyshev system (see, e.g., Krein and Nudelman, 1977, Chapter II for the formal definition of such systems). This property implies that some key determinants do not vanish, and any non-zero "exponential polynomial" $v \mapsto \sum_{k=1}^{K} \alpha_k \exp(\zeta_k v)$ does not have more than $K - 1$ zeros.

We now turn to necessary conditions for identification. The following result is a partial converse of Lemma 1.2.4 and Theorem 1.2.5 above.

Theorem 1.2.6 *Suppose that Assumptions 1.2.1-1.2.2 hold and $T = \tau + 1 \geq 3$. Then:*

1. *If $\mathbb{P}(|\{X_1, \ldots, X_T\}| = T) = 0$, then $B = \mathbb{R}^{K*}$.*
2. *If $T = 3$, then, for any $b \in R$, there exists a distribution of $\gamma|X$ such that for the corresponding distribution of $Y|X$, $b \in B$.*[7]

The first result shows that for our conditional moments to have any identifying power, there must exist trajectories of $X = (X_1, \ldots, X_T)$ with distinct values at all

[7] Note that B depends on the distribution of $\gamma|X$ but as before, we leave this dependence implicit.

periods. Since we focus here on $T \geq 3$, this excludes in particular the case where X_t is binary. More generally, if all components of X_t are binary, one must have $K > \log(T)/\log(2)$ for our moment conditions to have some identifying power. The second result shows that when $T = 3$, one cannot improve (1.2.5), at least in a uniform sense over conditional distributions of γ. Specifically, for any $b \in R$, there exists a data generating process satisfying Assumptions 1.2.1-1.2.2 and for which $b \in B$. Note however that failure of point identification at b implies strong restrictions on the distribution of $\gamma|X$. If $b \in B$ with $b \neq \beta_0$, then, for almost all x,

$$\mathbb{E}\left[a_1(\gamma, x)D_1(x, b) + a_2(\gamma, x)D_2(x, b)|X = x\right] = 0, \tag{1.2.6}$$

where $a_i(\gamma, x)$ is defined in (1.6.20). Namely, the distribution of $\gamma|X$ should satisfy a conditional moment restriction (note that (1.2.6) trivially holds when b is replaced by β_0, because $D_1(x, \beta_0) = D_2(x, \beta_0) = 0$). A violation of (1.2.6) on a set of x of positive measure is sufficient to discard b from B.[8]

1.3 GMM Estimation

1.3.1 Efficiency Bounds

We now suppose point identification based on (1.2.4) (namely, $B = \{\beta_0\}$) and discuss estimation of β_0. Let $R(X) = \mathbb{E}[\nabla_\beta m(Y, X; \beta_0)|X]$, $\Omega(X) = \mathbb{V}[m(Y, X; \beta_0)|X]$ (so that $\Omega(X) \in \mathbb{R}$) and define, provided that it exists,

$$V_0 := \mathbb{E}\left[\Omega(X)^{-1}R(X)R(X)'\right]^{-1}.$$

As shown by Chamberlain (1987), asymptotically optimal estimators of β_0 based on (1.2.4) have an asymptotic variance equal to V_0. The standard way to construct such estimators consists in two steps: first, one uses the unconditional moment $g(X)m(Y, X; \beta)$ for some $g(\cdot)$ and second, one estimates the optimal instruments $g^\star(X) := R(X)/\Omega(X)$. Such estimators, however, are not consistent if

$$\mathbb{E}[g(X)m(Y, X; \beta)] = 0 \text{ or } \mathbb{E}[g^\star(X)m(Y, X; \beta)] = 0$$

for $\beta \neq \beta_0$; see Dominguez and Lobato (2004). Instead, we can use an efficient GMM estimator exploiting the continuum of moment conditions associated with (1.2.4). We refer in particular to Section 4 in Hsu and Kuan (2011) and Section 2.5 in Lavergne and Patilea (2013) for the construction of such estimators.

These GMM estimators are optimal among those based on (1.2.4). However, it is not obvious that (1.2.4) actually exhausts all the possible restrictions induced by the model, and therefore that V_0 is the semiparametric efficiency bound of β_0. Theorem 1.3.1 below shows that this is the case for $T = \tau+1 = 3$ under the following conditions.

[8] Related to this, we establish point identification of β_0 under some restrictions on the conditional distribution of $\gamma|X$ in a previous version of Davezies et al. (2020).

Assumption 1.3.1 *1. There exists $t \in \{1, \dots, T\}$ such that $\mathbb{E}[X_t X_t']$ is nonsingular.*

2. $\mathbb{E}\left[\Omega^{-1}(X)R(X)R(X)'\right]$ exists and is nonsingular.

3. $|Supp(\gamma|X)| \geq 10$ almost surely.

The first condition is a mild restriction on X. The second condition is a local identifiability condition, which is neither weaker nor stronger than $B = \{\beta_0\}$. The third condition is weaker than that imposed by Chamberlain (2010), namely $\text{Supp}(\gamma|X) = \mathbb{R}$. Intuitively, if $\gamma|X$ has few points of support, moments of $\gamma|X$ are restricted, and we may exploit this to produce additional restrictions that would improve an estimation of β_0 based solely on (1.2.4).

Theorem 1.3.1 *Assume $T = \tau + 1 = 3$ with $\lambda_2 \neq 2$ and Assumptions 1.2.1, 1.2.2 and 1.3.1 hold. Then the semiparametric efficiency bound of β_0, $V^\star(\beta_0)$, is finite and satisfies $V^\star(\beta_0) = V_0$.*

Intuitively, this result states that all the information content of the model is included in the conditional moment restriction $\mathbb{E}[m(Y, X; \beta_0)|X] = 0$. It complements, for $T = \tau + 1 = 3$, the result of Hahn (1997), which states that the conditional maximum likelihood estimator is the efficient estimator of β_0 if F is logistic. Note however that we cannot compare his bound with ours in the logistic case: for this distribution, $w_2 = 0$, and for identification reasons, this case is excluded from our family of generalized logistic distributions with $\tau = 2$. We refer to Footnote 4 above for more details about this.

1.3.2 Unbalanced Panel

In many applications, as that considered below, panel data are unbalanced. To handle this case, we can simply consider, for each individual, all possible subsets of periods of size $\tau + 1$ and form the corresponding moment conditions. Specifically, suppose that the set of periods available for individual i is $\mathcal{T}_i \subset \{1, \dots, T\}$. Thus, we observe the sample $((Y_{it}, X_{it})_{t\in\mathcal{T}_i})_{i=1,\dots,n}$. Let us assume that the selection of periods is (conditionally) exogenous, namely

$$\mathcal{T}_i \perp\!\!\!\perp (Y_{it})_{t\geq 1} | (X_{it})_{t\geq 1}, \gamma_i. \tag{1.3.1}$$

Then, we basically get back to the case $T = \tau + 1$ by considering the moment vector

$$\psi(Y_i, X_i, \mathcal{T}_i, \beta) = \mathbb{1}\{|\mathcal{T}_i| \geq \tau + 1\} \sum_{\substack{t_1<\dots<t_{\tau+1} \\ (t_1,\dots,t_{\tau+1})\in\mathcal{T}_i^{\tau+1}}} g(X_{it_1}, \dots, X_{it_{\tau+1}}) m((Y_{it}, X_{it})_{t\in\{t_1,\dots,t_{\tau+1}\}}, \beta).$$

for some function $g(X_{it_1}, \ldots, X_{it_{\tau+1}}) \in \mathbb{R}^L$, with $L \geq K$. Condition (1.3.1) ensures that $\mathbb{E}\left[\psi(Y_i, X_i, \mathcal{T}_i, \beta_0)\right] = 0$. Then, we can consider the GMM estimator

$$\widehat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \psi(Y_i, X_i, \mathcal{T}_i, \beta) \right)' \widehat{W} \left(\sum_{i=1}^{n} \psi(Y_i, X_i, \mathcal{T}_i, \beta) \right), \tag{1.3.2}$$

for some symmetric positive definite $\widehat{W}$. This idea also applies to balanced panel data for which $T > \tau + 1$. In such a case, $\mathcal{T}_i = \{1, \ldots, \tau + 1\}$ and (1.3.1) automatically holds.

1.4 Application to Brender and Drazen (2008)

Brender and Drazen (2008) study how budget deficits and economic growth affect reelection. To this end, they gather data from multiple sources on 74 countries, over the period 1960-2003. They use two definitions for their binary outcome variable REELECT, one where reelection is defined in a "narrow" sense and another where it is "expanded", following here their terminology. This also leads to two different samples, as REELECT may be missing in the narrow sense but equal to 0 in the expanded sense. The covariates related to budget deficits are BALCH_term and BALCH_ey. BALCH_term corresponds to the change in ratio of the central government's balance to GDP over the term in office. BALCH_ey is the change in the balance/GDP ratio between the year preceding the election and the election year. The variable GDPPC_gr is the average annual growth rate of real GDP per capita between two election years. The authors also include in their models two controls, namely a dummy for a new democracy and a dummy of having a majoritarian electoral system. We refer to Brender and Drazen (2008) for more details about the data.

In their main specification, Brender and Drazen (2008) consider a simple logit model, see Table 2 therein. Then, as a robustness check (see their Table 3), they estimate a fixed effect logit model. They show that their main results are robust to including fixed effects. However, the assumption of logistic errors is not warranted, so we investigate whether the results are robust to this assumption, by considering instead the family of generalized logistic distribution, with $\tau = 2$. We focus on the sample of developed countries as the sample of less developped countries is very small, and thus leads to noisy estimates. Note that the data are not balanced at all: some countries are only observed for 4 periods in the narrow sample (resp. 5 in the expanded sample), while others are observed over 13 (resp. 14) periods. We thus apply the procedure mentioned in Section 1.3.2. The vector of instruments $g(X_{it_1}, X_{it_2}, X_{it_3})$ is simply the list of the corresponding 15 variables (as $X_{it} \in \mathbb{R}^5$), demeaned over these three periods. We consider $\lambda_2 = 1.2, 1.4, 1.6$ and 1.8. We do not consider larger values of λ_2 as they seem to lead to numerical instabilities.[9] Finally, as the GMM objective function may have local optima, we consider 200 random initial

[9] This may be because $|M_t(x; \beta)|$ increases quickly with λ_2, due to the exponential function.

points and pick the vector of parameters minimizing the corresponding final objective function.

The results are presented in Table 1.1. Because the coefficients themselves are not comparable, we focus on the sign of BALCH_ey and on the relative effects with respect to BALCH_ey; note that we were able to recover the exact same estimates as Brender and Drazen (2008) in their Tables 2 and 3. We choose BALCH_ey as the reference variable for relative effects because its coefficient should not be 0, and it has the largest t-test on the logit and fixed effect logit model. For the three methods, the t-statistics of relative effects under the null hypothesis are obtained using the estimated asymptotic variance of $\widehat{\beta}$.

Overall, at least two important results seem robut to the distributional assumption on the unobserved terms. First, the sign of BALCH_ey is always positive. Second, the relative effect of BALCH_term and BALCH_ey remain quite stable when considering our FE generalized logistic model, with fluctuations between 0.27 and 0.52 depending on the sample and value of λ_2 that we consider. At the 10% level, we cannot reject that the effect of BALCH_term is actually 0, except in the narrow sample with $\lambda_2 = 1.8$. But the test was already close to not being rejected with the FE logit model on the narrow sample (p-value=0.097), and not rejected with the simple and FE logit models based on the expanded sample (p-values=0.124 and 0.204 respectively). So the most important results seem overall robust to the change of specification we consider. Other results fluctuate slightly more: the fact of being a new democracy had a positive and borderline significant effect with the expanded sample (p-value=0.099). It is not significant anymore with our model, the coefficient being sometimes even negative.

TABLE 1.1: Estimates of Relative Effects of Budget Balances and Growth on the Probability of Reelection in Developed Economies

	Logit	FE logit	FE generalized logit			
λ_2			1.2	1.4	1.6	1.8
Narrow sample						
Sign of BALCH_ey	>0	>0	>0	>0	>0	>0
BALCH_term/BALCH_ey	0.54	0.55	0.48	0.37	0.37	0.52
	(2.34)	(1.82)	(0.03)	(0.06)	(0.03)	(2.49)
GDPPC_gr/BALCH_ey	-0.04	0.27	-0.35	-0.34	-0.33	-0.37
	(0.17)	(0.70)	(0.04)	(0.07)	(0.04)	(1.04)
New democraties/BALCH_ey	0.03	0.07	0.05	0.05	0.05	-0.20
	(2.69)	(1.62)	(0.04)	(0.08)	(0.04)	(0.03)
Majoritarian	0.02	0.07	-0.10	0.00	0.00	-0.03
electoral system/BALCH_ey	(1.31)	(1.52)	(0.03)	(0.02)	(0.01)	(0.20)
Expanded sample						
Sign of BALCH_ey	>0	>0	>0	>0	>0	>0
BALCH_term/BALCH_ey	0.40	0.36	0.34	0.27	0.37	0.47
	(1.44)	(1.35)	(0.85)	(1.64)	(0.72)	(0.42)
GDPPC_gr/BALCH_ey	0.09	0.46	-0.09	0.00	-0.14	-0.28
	(0.30)	(1.20)	(0.22)	(0.00)	(0.29)	(0.61)
New democraties/BALCH_ey	0.04	0.09	0.02	0.01	-0.00	-0.15
	(3.11)	(1.81)	(0.32)	(0.08)	(0.00)	(0.26)
Majoritarian	0.02	0.04	-0.15	-0.10	-0.21	-0.67
electoral system/BALCH_ey	(1.74)	(1.12)	(0.99)	(1.54)	(1.08)	(1.14)

Notes: Analytical t-statistics of the coefficient ratios are under parentheses. The estimated asymptotic variance of the simple logit model is obtained through clustering at the country level. Both samples include 23 countries, with on average 7.1 (resp. 7.8) periods per country in the narrow (resp. expanded) sample.

1.5 Conclusion

This Chapter studies the identification and root-n estimation of the common slope parameter in a static panel binary model with exogenous and bounded regressors. We first show that when $T \geq 3$ and the unobserved terms belong to a family of generalized logistic distribution, a conditional moment restriction holds. Then, we study the identified set corresponding to these restrictions. In particular, under a restriction on the distribution of covariates only, relative effects are point identified, no matter the distribution of the individual effect. Our identification results lead to a GMM estimator that reaches the semiparametric efficiency bound when $T = 3$. Estimating this model may serve as a robustness check for the fixed effect logit model, something we illustrate in the application.

This Chapter also leaves a few questions unanswered. A first one is whether the family of F considered here is the only one for which point identification can be achieved (see Section 1.7 below). Another one is whether the GMM estimator still reaches the semiparametric efficiency bound when $T > 3$. Both questions raise difficult issues and deserve future investigation.

1.6 Proofs of the Results

For any real $a \in \mathbb{R}$, we let $\text{sgn}(a) := \mathbb{1}(a > 0) - \mathbb{1}(a < 0)$. For any subset A of a reference space E, we let A^c denote the complement of A in E. The following lemma on "exponential polynomials" is key in the proof of Theorems 1.2.5 and 1.3.1.

Lemma 1.6.1 *Let $K \geq 1$, $(\zeta_1, \ldots, \zeta_K)$ be K distinct real numbers, $(\alpha_1, \ldots, \alpha_K)' \in \mathbb{R}^K$, $(\alpha_1, \ldots, \alpha_K) \neq (0, \ldots, 0)$ and $P(x) := \sum_{k=1}^K \alpha_k \exp(\zeta_k x)$. Then P has at most $K - 1$ distinct roots.*

The proof is by induction on K and Rolle's theorem, see e.g. Chapter 2, section 2 of Krein and Nudelman (1977).

1.6.1 Proposition 1.2.1

The sufficient part is obvious. To prove necessity, suppose $\beta_0 \neq 0$. Since $\mathbb{E}[(X_t - X_{t'})(X_t - X_{t'})']$ is non singular, there exists a subset $\mathcal{S}$ of the support of $(X_t, X_{t'})$ such that $\mathbb{P}(\mathcal{S}) > 0$ and for all $(x_t, x_{t'}) \in \mathcal{S}$, $(x_t - x_{t'})'\beta_0$ has constant, non-zero sign. Without loss of generality let us assume $(x_t - x_{t'})'\beta_0 > 0$. Let $G(u) := F(u)/(1 - F(u))$. Because G is strictly increasing, we have, for all $g \in \mathbb{R}$,

$$G(x_t'\beta_0 + g) > G(x_{t'}'\beta_0 + g).$$

Equivalently,

$$F(x_t'\beta_0 + g)(1 - F(x_{t'}'\beta_0 + g)) > F(x_{t'}'\beta_0 + g)(1 - F(x_t'\beta_0 + g)).$$

In other words,

$$\mathbb{P}(Y_1 = 1, Y_{t'} = 0 | X_t = x_t, X_{t'} = x_t', \gamma = g) > \mathbb{P}(Y_1 = 0, Y_{t'} = 1 | X_t = x_t, X_{t'} = x_t', \gamma = g),$$

and the result follows by integration over g.

1.6.2 Theorem 1.2.3

Let us define

$$A(x,\gamma;\beta) := \begin{pmatrix} \sum_{j=1}^{T-1} w_j \exp(\lambda_j(x_1'\beta+\gamma)) & \dots & \sum_{j=1}^{T-1} w_j \exp(\lambda_j(x_T'\beta+\gamma)) \\ \exp(\lambda_1 x_1'\beta) & \dots & \exp(\lambda_1 x_T'\beta) \\ \vdots & & \vdots \\ \exp(\lambda_{T-1} x_1'\beta) & \dots & \exp(\lambda_{T-1} x_T'\beta) \end{pmatrix}.$$

Let $A_i(x,\gamma;\beta)$ denote the ith row of $A(x,\gamma;\beta)$. Then

$$A_1(x,\gamma;\beta) = \sum_{j=1}^{T-1} w_j \exp(\lambda_j\gamma) A_{j+1}(x,\gamma;\beta).$$

It follows that for all $(x,\gamma) \in \text{Supp}(X) \times \mathbb{R}$,

$$\det A(x,\gamma;\beta_0) = 0.$$

By Assumption 1.2.2 and since we focus on the first type therein, we have $G(u) := F(u)/(1-F(u)) = \sum_{j=1}^{T-1} w_j \exp(\lambda_j u)$. Now, developing $\det A(x,\gamma;\beta_0)$ with respect to the first row yields, by definition of the function m,

$$\sum_{y\in\{0,1\}^T} m(y,x;\beta_0) \prod_{t:y_t=1} G(x_t'\beta_0+\gamma) = 0.$$

Multiplying this equality by $\prod_t (1-F(x_t'\beta_0+\gamma))$ we obtain

$$\sum_{y\in\{0,1\}^T} \left[m\left(y,x;\beta_0\right) \prod_{t:y_t=1} F(x_t'\beta_0+\gamma) \prod_{t:y_t=0} (1-F(x_t'\beta_0+\gamma)) \right] = 0.$$

This equation is equivalent to $\mathbb{E}\left[m(Y,X;\beta_0)|X,\gamma\right] = 0$ a.s. The result follows.

1.6.3 Lemma 1.2.4

Let $b \in \widetilde{B}^c$ and let us prove that $b \notin B$. Fix $x \in \mathcal{D}(b)$ and let $\mathcal{J}_x := \{j \in \{1,\dots,T-1\} : D_j(x;b) \neq 0\}$ and

$$a_j(x) := w_j \mathbb{E}\left[\frac{\exp(\lambda_j\gamma)}{\prod_{t=1}^T \left(1+\sum_{k=1}^{T-1} w_k \exp(\lambda_k(x_t'\beta_0+\gamma))\right)} \Big| X = x \right]. \tag{1.6.1}$$

Then $\mathcal{J}_x \neq \emptyset$ and

$$\mathbb{E}[m(Y,X;b)|X=x] = \sum_{j\in\mathcal{J}_x} a_j(x) D_j(x;b). \tag{1.6.2}$$

Moreover, $a_j(x) > 0$ and all the $D_j(x;b)$ for $j \in \mathcal{J}_x$ have the same sign. Thus, $\mathbb{E}[m(Y,X;b)|X=x] \neq 0$. Because $b \in \widetilde{B}^c$, we have, by definition of $\widetilde{B}$, $\mathbb{P}(X \in$

$\mathcal{D}(b)) > 0$. Thus, $\mathbb{E}[m(Y, X; b)|X = x] \neq 0$ with positive probability, implying $b \notin B$.

1.6.4 Theorem 1.2.5

Part 1

a. $B_k \subset R_k := \{c\beta_{0k} : c \in \{0\} \cup (1/\lambda_{T-1}, \lambda_{T-1})\}$.

Let us fix $k \in \{1, \ldots, K\}$, $b = (b_1, \ldots, b_K)$ and define

$$\mathcal{X}_{0k} := \{x \in \text{Supp}(X) : x_{j,1} = \cdots = x_{j,T} = 0 \; \forall j \neq k, |\{x_{k,1}, \ldots, x_{k,T}\}| = T\}.$$

First, suppose that $\beta_{0k} = 0$ and $b_k \neq 0$. Then, $D_j(x; b)$ does not depend on j. Moreover, because

$$|\{x_1'b, \ldots, x_T'b\}| = |\{x_{k,1}b_k, \ldots, x_{k,T}b_k\}| = T,$$

we have $D_j(x; b) \neq 0$ by properties of Chebyshev systems. Thus, $x \in \mathcal{D}(b)$, implying that $\mathcal{X}_{0k} \subset \mathcal{D}(b)$. By Assumption 1.2.3, $\mathbb{P}(X \in \mathcal{X}_{0k}) > 0$. Hence, $\mathbb{P}(X \in \mathcal{D}(b)) > 0$. By Lemma 1.2.4, $b_k \notin B_k$ and $B_k \subset \{0\} = R_k$.

Now, suppose $\beta_{0k} \neq 0$. Then any $b_k \in \mathbb{R}$ can be written as $c\beta_{0k}$. We prove that if $c \notin \{0\} \cup (1/\lambda_{T-1}, \lambda_{T-1})$, then $\mathcal{X}_{0k} \subset \mathcal{D}(b)$. By Lemma 1.2.4 again, this shows that $B_k \subset R_k$. Let us first suppose that $c \notin \{1/\lambda_{T-1}, \lambda_{T-1}\}$ and fix $x \in \mathcal{X}_{0k}$. Let us show that for each $(j, j') \in \{1, \ldots, T-1\}^2$,

$$\text{sgn}(D_j(x; b)) = \text{sgn}(D_{j'}(x; b)) \neq 0. \tag{1.6.3}$$

If $c \in (-\infty, 0)$, we have

$$c\lambda_{T-1} < \ldots < c\lambda_2 < c < 0 < 1 < \lambda_2 < \ldots < \lambda_{T-1}. \tag{1.6.4}$$

If $c \in (0, 1/\lambda_{T-1})$, we have

$$0 < c < c\lambda_2 < \ldots < c\lambda_{T-1} < 1 < \lambda_2 < \ldots < \lambda_{T-1}. \tag{1.6.5}$$

Else, $c \in (\lambda_{T-1}, +\infty)$ and we have

$$1 < \lambda_1 < \ldots < \lambda_{T-1} < c < c\lambda_2 < \ldots < c\lambda_{T-1}. \tag{1.6.6}$$

Let p_j denote the number of transpositions (ie permutations exchanging two elements, leaving the others fixed) needed to sort $\tilde{\lambda}^j := (\lambda_j, c, c\lambda_2, \ldots, c\lambda_{T-1})'$ in ascending order. It is clear from Equations (1.6.4)-(1.6.6) that $p_j = p_{j'} = p$ for all $(j, j') \in$

$\{1,\dots,T-1\}^2$. Let $\tilde{\lambda}^{sj}$ denote the sorted version of $\tilde{\lambda}^j$ and define

$$D_j(x;b,\lambda) =: \det \begin{pmatrix} \exp(\lambda_j x_1'\beta_0) & \dots & \exp(\lambda_j x_T'\beta_0) \\ \exp(\lambda_1 x_1' b) & \dots & \exp(\lambda_1 x_T' b) \\ \vdots & & \\ \exp(\lambda_{T-1} x_1' b) & \dots & \exp(\lambda_{T-1} x_T' b) \end{pmatrix},$$

so that $D_j(x;b) = D_j(x;b,\lambda)$. Because $x \in \mathcal{X}_{0k}$, we have

$$D_j(x;b,\lambda) = \det \begin{pmatrix} \exp(\lambda_j x_{k,1}\beta_{0k}) & \dots & \exp(\lambda_j x_{k,T}\beta_{0k}) \\ \exp(c\lambda_1 x_{k,1}\beta_{0k}) & \dots & \exp(c\lambda_1 x_{k,T}\beta_{0k}) \\ \vdots & & \\ \exp(c\lambda_{T-1} x_{k,1}\beta_{0k}) & \dots & \exp(c\lambda_{T-1} x_{k,T}\beta_{0k}) \end{pmatrix} = D_j(x;\beta_0,\tilde{\lambda}^j).$$

Hence, for all $j \in \{1,\dots,T-1\}$,

$$\operatorname{sgn}\left(D_j(x;b)\right) = \operatorname{sgn}\left(D_j(x;\beta_0,\tilde{\lambda}^j)\right) = (-1)^p \operatorname{sgn}\left(D_j(x;\beta_0,\tilde{\lambda}^{sj})\right).$$

Now, let $\overline{p}$ be the number of pairwise coordinates permutations needed to sort the vector $(x_1'\beta_0,\dots,x_T'\beta_0)'$ in ascending order, and let x^s denote a rearrangement of x such that $x_1^{s\prime}\beta_0 < \dots < x_T^{s\prime}\beta_0$. It follows that, for all $j \in \{1,\dots,T-1\}$,

$$\begin{aligned} \operatorname{sgn}\left(D_j(x;b)\right) &= (-1)^p \operatorname{sgn}\left(D_j(x;\beta_0,\tilde{\lambda}^{sj})\right) \\ &= (-1)^{p+\overline{p}} \operatorname{sgn}\left(D_j(x^s;\beta_0,\tilde{\lambda}^{sj})\right) \\ &= (-1)^{p+\overline{p}}, \end{aligned}$$

where the last equality follows by properties of Chebyshev systems. The last equality implies that (1.6.3) holds. Hence $x \in \mathcal{D}(b)$, implying $\mathbb{P}(X \in \mathcal{D}(b)) > 0$.

Finally, consider the case where $b = c\beta_0$ with $c \in \{1/\lambda_{T-1}, \lambda_{T-1}\}$. By continuity of the determinant and (1.6.3), we either have $0 \le \min_{j=1,\dots,T-1} D_j(x;b) \le \max_{j=1,\dots,T-1} D_j(x;b)$ or $0 \ge \max_{j=1,\dots,T-1} D_j(x;b) \ge \min_{j=1,\dots,T-1} D_j(x;b)$. Moreover, $D_{T-1}(x;\beta_0/\lambda_{T-1}) \neq 0$ and $D_1(x;\lambda_{T-1}\beta_0) \neq 0$. Therefore, whatever the value of c ($1/\lambda_{T-1}$ or λ_{T-1}), we have $x \in \mathcal{D}(b)$. Then, again, $\mathbb{P}(X \in \mathcal{D}(b)) > 0$. The result follows.

b. $|B| < \infty$**.**

Because $|B| \le \prod_{k=1}^K |B_k|$, it suffices to prove that for each k, $|B_k| < \infty$. Fix k. If $\beta_{0k} = 0$, then $B_k = \{0\}$ and we have nothing to prove. Otherwise, let $b = (b_1,\dots,b_K) \in B$ and fix $x = (x_1,\dots,x_T) \in \mathcal{X}_{0k}$. Let $c \in \{0\} \cup (1/\lambda_{T-1}, \lambda_{T-1})$ be such that $b_k = c\beta_{0k}$. By Equation (1.6.2), we have $\sum_{j=1}^{T-1} a_j(x) D_j(x;b) = 0$, where $a_j(x)$ is defined by (1.6.1). Moreover, by definition of $\mathcal{X}_{0k}$, we have $D_j(x;b) = D_j(x;c\beta_0)$.

Then, c satisfies

$$\sum_{j=1}^{T-1} a_j(x) D_j(x; c\beta_0) = 0, \tag{1.6.7}$$

Developing $D_j(x; c\beta_0)$ with respect to the first line, and using the definition of the determinant, we obtain

$$\sum_{t=1}^{T} (-1)^{t+1} \sum_{j=1}^{T-1} a_j(x) \exp(\lambda_j x_t' \beta_0) \sum_{\sigma \in \mathfrak{S}_t} \varepsilon(\sigma) \exp\left[\left(\sum_{s \neq t} \lambda_{\sigma(s)} x_s' \beta_0\right) c\right] = 0, \tag{1.6.8}$$

where $\mathfrak{S}_t$ is the set of bijections from $\{1, \ldots, T\} \backslash \{t\}$ to $\{1, \ldots, T-1\}$ and $\varepsilon(\sigma)$ denotes the parity of σ (we can assimilate σ to a permutation by assimilating $\{1, \ldots, T\} \backslash \{t\}$ with $\{1, \ldots, T-1\}$, keeping the natural ordering of both sets). The left-hand side of (1.6.8) is a function of c of the form $\sum_{k=1}^{K} d_k \exp(b_k c)$, with $K \leq T!$ (the inequality arises because some coefficients in the exponential monomials may be equal). Let us show that $d_k \neq 0$ for at least one k. First, remark that $x_t' \beta_0 = x_{k,t} \beta_{0,k}$. Then, because $|\{x_{k,1}, \ldots, x_{k,T}\}| = T$ and $\beta_{0k} \neq 0$, we can assume without loss of generality, up to a rearrangement of periods, that $x_1' \beta_0 < \ldots < x_T' \beta_0$. Let I_t be the element of $\mathfrak{S}_t$ such that $I_t(s) = s - \mathbb{1}\{s \geq t+1\}$. By the rearrangement inequality, for all $\sigma \in \mathfrak{S}_t \backslash \{I_t\}$,

$$\sum_{s \neq t} \lambda_{\sigma(s)} x_s' \beta_0 < \sum_{s \neq t} \lambda_{I_t(s)} x_s' \beta_0.$$

Moreover, for all $t \in \{1, \ldots, T\}$, let

$$g(t) := \sum_{s \neq t} \lambda_{I_t(s)} x_s' \beta_0.$$

Because $I_t(s) = I_{t-1}(s)$ for all $t > 1$ and $s \leq t-2$ or $s \geq t+1$, we have

$$g(t) - g(t-1) = \lambda_{t-1} x_{t-1}' \beta_0 - \lambda_{t-1} x_t' \beta_0 < 0.$$

Hence, the exponential monomial with highest coefficient in (1.6.8) is

$$\exp\left[\left(\sum_{s \neq 1} \lambda_{s-1} x_s' \beta_0\right) c\right]$$

and we can obtain it only by letting $t = 1$ and $\sigma = I_1$. Because $a_j(x) > 0$ for all j, the coefficient of this monomial is $\sum_{j=1}^{T-1} a_j(x) \exp(\lambda_j x_1' \beta_0) > 0$. Therefore, at least one d_k in the exponential polynomial $\sum_{k=1}^{K} d_k \exp(b_k c)$ satisfies $d_k \neq 0$. Then, by Lemma 1.6.1, the equation $\sum_{k=1}^{K} d_k \exp(b_k c) = 0$ has at most $T! - 1$ solutions. Thus, $|B_k| \leq T! - 1$. The result follows.

Part 2

The point identification of relative marginal effects is obvious given the other results, which we prove in turn.

a. Equation (1.2.5) **holds.**

Let us define, for all $b \in \mathbb{R}^{K*}$,

$$\mathcal{X}_1(b) = \Big\{ x = (x_1, \dots, x_T) \in \text{Supp}(X) : \exists (s,t) \in \{1,\dots,T\}^2 : x_s'b = x_t'b,\ x_s'\beta_0 \neq x_t'\beta_0, \\ \text{and } \forall (s',t') \in \{1,\dots,T\}^2, s' \neq t', \{s',t'\} \neq \{s,t\} :\ x_{s'}'b \neq x_{t'}'b \Big\}.$$

The proof of is divided into three steps. First, we prove that $\mathcal{X}_1(b) \subset \mathcal{D}(b)$, for all $b \in \mathbb{R}^{K*} \backslash \text{lin}(\beta_0)$. In a second step, we prove that $\widetilde{B} \subset \text{lin}(\beta_0)$. Finally, the third step shows that $\widetilde{B} \subset R$.

First step: $\mathcal{X}_1(b) \subset \mathcal{D}(b)$ for all $b \in \mathbb{R}^{K*} \backslash lin(\beta_0)$.

Let $x \in \mathcal{X}_1(b)$ and (s,t) be as in the definition of $\mathcal{X}_1(b)$. Developing $D_j(x;b)$ according to the first row, we obtain, for all $j \in \{1,\dots,T-1\}$,

$$D_j(x;b) = \sum_{\ell=1}^{T} (-1)^{\ell+1} \exp(\lambda_j x_\ell' \beta_0) D_j^{-\{1,\ell\}}(x;b),$$

where $D_j^{-\{1,\ell\}}(x;b)$ denotes the determinant of the matrix in $D_j(x;b)$ once its first row and ℓth column have been removed. Remark that, for all $j \in \{1,\dots,T-1\}$, for all $\ell \in \{1,\dots,T\} \backslash \{s,t\}$, $D_j^{-\{1,\ell\}}(x;b) = 0$, and

$$D_j^{-\{1,t\}}(x;b) = (-1)^{|s-t|-1} D_j^{-\{1,s\}}(x;b).$$

As a result,

$$\begin{aligned} D_j(x;b) &= (-1)^{s+1} \exp(\lambda_j x_s' \beta_0) D_j^{-\{1,s\}}(x;b) + (-1)^{t+1} \exp(\lambda_j x_t' \beta_0) D_j^{-\{1,t\}}(x;b) \\ &= D_j^{-\{1,s\}}(x;b) \left[(-1)^{s+1} \exp(\lambda_j x_s' \beta_0) + (-1)^{t+1} \exp(\lambda_j x_t' \beta_0)(-1)^{|s-t|-1} \right] \\ &= D_j^{-\{1,s\}}(x;b)(-1)^{s+1} \left[\exp(\lambda_j x_s' \beta_0) - \exp(\lambda_j x_t' \beta_0) \right], \end{aligned}$$

where we have used $(-1)^{|s-t|+t} = (-1)^s$. Now, $D_j^{-\{1,s\}}(x;b)$ does not depend on j and by definition of Chebyshev systems, $D_j^{-\{1,s\}}(x;b) \neq 0$. Also, the sign of the term inside brackets is equal to the sign of $(x_s - x_t)'\beta_0$, and thus does not depend on j. Hence for all $(j,j') \in \{1,\dots,T-1\}^2$,

$$\text{sgn}\left(D_j(x;b)\right) = \text{sgn}\left(D_{j'}(x;b)\right) \neq 0,$$

which shows that $x \in \mathcal{D}(b)$.

Second step: $\widetilde{B} \subset lin(\beta_0)$.

Fix $b \notin \text{lin}(\beta_0)$, $b \neq 0$ and let us prove that $\mathbb{P}(X \in \mathcal{D}(b)) > 0$. The result will then follow by Lemma 1.2.4.

Suppose without loss of generality that (s,t) in Assumption 1.2.4 is equal to $(1,2)$. By that assumption, there exists $\widetilde{x} := (x', x', x_3', \ldots, x_T')' \in \text{Supp}(X)$ and a neighborhood $\widetilde{V}$ of $\widetilde{x}$ included in $\text{Supp}(X)$. Since b and β_0 are not collinear, there exists $(u_1', u_2')' \in \mathbb{R}^{2K}$ such that $(u_1 - u_2)'b = 0$ and $(u_1 - u_2)'\beta_0 \neq 0$. Moreover, up to replacing $(u_1', u_2')'$ by $c(u_1', u_2')'$ with $c \neq 0$, $(u_1', u_2')'$ can be chosen of arbitrarily small norm.

Now, let $x_1 = x_2 = x$ and

$$\mathcal{A}(u_1, u_2) = \Big\{(u_3', \ldots, u_T')' \in \mathbb{R}^{K(T-2)} : \forall (s,t) \in \{1, \ldots, T\}^2, s \neq t, \{s,t\} \neq \{1,2\} : \\ (u_s - u_t + x_s - x_t)'b \neq 0\Big\},$$

The set $\mathcal{A}(u_1, u_2)$ is dense as the intersection of open, dense subsets of $\mathbb{R}^{K(T-2)}$. Hence, there exists $(u_3', \ldots, u_T')' \in \mathcal{A}(u_1, u_2)$ with arbitrarily small norm. Then, we can ensure that $u := (u_1', \ldots, u_T')'$ satisfies $x^* := \widetilde{x} + u \in \widetilde{V}$. Moreover, by construction, $x^* \in \mathcal{X}_1(b)$. Then, Step 1 implies $x^* \in \mathcal{D}(b)$ and $D_j(x; b) \neq 0$ for all j. By continuity of the map $x \mapsto D_j(x; b)$ and Assumption 1.2.4, there exists a neighborhood of x^*, $\mathcal{V} \subset \mathcal{D}(b)$ such that $\mathbb{P}(X \in \mathcal{V}) > 0$. Hence, $P(X \in \mathcal{D}(b)) > 0$.

Third step: $\widetilde{B} \subset R$.

We just have to prove that if $b = c\beta_0$ with $c \in (-\infty, 1/\lambda_{T-1}] \cup [\lambda_{T-1}, +\infty)$ and $c \neq 0$ (since $\beta_0 \neq 0$), then $b \notin B$. The reasoning is exactly the same as in Part 1.a, with just one change: Instead of considering $x \in \mathcal{X}_{0k}$, we consider $x \in \mathcal{X}_0$, with

$$\mathcal{X}_0 := \{x \in \text{Supp}(X) : \left|\{x_1'\beta_0, \ldots, x_T'\beta_0\}\right| = T\}.$$

b. $|B| \leq T! - 1$**.**

The reasoning is exactly the same as in Part 1.b, with just two changes. First, we reason directly on B, not on B_k. Second, instead of considering $x \in \mathcal{X}_{0k}$, we consider $x \in \mathcal{X}_0$.

c. $|B| \leq 2$ **when** $T = 3$.

For any $b = c\beta_0 \in B$, we have, as in Eq. (1.6.7),

$$a_1(x)D_1(x; c\beta_0) + a_2(x)D_2(x; c\beta_0) = 0 \tag{1.6.9}$$

for almost all $x \in \text{Supp}(X)$. Suppose there exist three distinct solutions $1, c_1, c_2$ to Equation (1.6.9), with $1/\lambda_2 < c_1 < c_2 < \lambda_2$. Multiply Eq. (1.6.9), evaluated at $c = c_1$, by $D_2(x; c_2\beta_0)$. Similarly, multiply Eq. (1.6.9), evaluated at $c = c_2$, by $D_2(x; c_1\beta_0)$. Substracting the two expressions, we obtain, since $a_1(x) > 0$,

$$D_1(x; c_1\beta_0)D_2(x; c_2\beta_0) - D_1(x; c_2\beta_0)D_2(x; c_1\beta_0) = 0. \tag{1.6.10}$$

For any $x \in \mathcal{X}_0$, let $u_t := x_t'\beta_0$. Fixing u_2 and u_3, (1.6.10) may be written as

$$P(u_1) := \sum_{k=1}^{13} \alpha_k \exp(\zeta_k u_1) = 0, \tag{1.6.11}$$

where the α_k and ζ_k are functions of (u_2, u_3). Suppose first that $c_2 > 1$. Some tedious algebra shows that the smallest ζ_k is $1 + c_1$, and its associated coefficient is equal to

$$\begin{aligned}\alpha_k =& \left[\exp(c_2(u_2 + \lambda_2 u_3)) - \exp(c_2(u_3 + \lambda_2 u_2))\right] \\ &\times \left[\exp(\lambda_2(u_2 + c_1 u_3)) - \exp(\lambda_2(u_3 + c_1 u_2))\right].\end{aligned}$$

Because $u_2 \neq u_3$ (as $x \in \mathcal{X}_0$), $\alpha_k \neq 0$. Hence P is nonzero and by Lemma 1.6.1, it has at most 12 zeros. However, under Assumption 1.2.4.2 and the second part of Assumption 1.2.4.3,

$$\left|\text{Supp}(X_1'\beta_0 | X_2'\beta_0 = u_2, X_3'\beta_0 = u_3) \backslash \{u_2, u_3\}\right| > 12.$$

Thus, in view of (1.6.11), P has strictly more than 12 zeros, a contradiction.

Second, suppose that $c_2 < 1$. Then, the largest ζ_k is $\lambda_2(1 + c_2)$, and its associated coefficient is equal to

$$\begin{aligned}\alpha_k =& -\left[\exp(u_2 + c_2 u_3)) - \exp(u_3 + c_2 u_2)\right] \\ &\times \left[\exp(c_1(u_2 + \lambda_2 u_3)) - \exp(c_1(u_3 + \lambda_2 u_2))\right].\end{aligned}$$

Again, $\alpha_k \neq 0$ and we reach a contradiction as before. The result follows.

1.6.5 Theorem 1.2.6

Part 1

Let us suppose that $\mathbb{P}(|\{X_1, \ldots, X_T\}| = T) = 0$. Let T_1 and $T_2 > T_1$ denote the two random dates, functions of X only, such that $X_{T_1} = X_{T_2}$ almost surely. For all $t \in \{1, \ldots, T\}$, let e_t denote the vector of $T - 1$ zeros and a 1 at coordinate t. Let $f(x; b) := \mathbb{E}\left[m(Y, X; b) | X = x\right]$. By definition,

$$f(X; b) = \sum_{y \in \{0,1\}^T} \mathbb{P}(Y = y|X) m(y, X; b). \tag{1.6.12}$$

Moreover, almost surely,

$$\begin{aligned}\mathbb{P}(Y = e_{T_1}|X) =& \int F(X_{T_1}'\beta_0 + \gamma) \prod_{t \neq T_1} (1 - F(X_t'\beta_0 + \gamma)) \mathrm{d}F_{\gamma|X}(\gamma) \\ =& \int F(X_{T_2}'\beta_0 + \gamma) \prod_{t \neq T_2} (1 - F(X_t'\beta_0 + \gamma)) \mathrm{d}F_{\gamma|X}(\gamma) \\ =& \mathbb{P}(Y = e_{T_2}|X).\end{aligned} \tag{1.6.13}$$

Next, remark that the matrices in $M_{T_1}(X;b)$ and $M_{T_2}(X;b)$ have the same columns but in different order, with $T_2 - T_1 - 1$ transpositions needed to obtain the same ordering. Thus, by definition of the determinant, $M_{T_1}(X;b) = -M_{T_2}(X;b)$, which implies

$$m(e_{T_1}, X; b) = -m(e_{T_2}, X; b). \tag{1.6.14}$$

Moreover, for all $s \notin \{T_1, T_2\}$, $m(e_s, X; b) = 0$ because M_s includes two identical columns (given that $X_{T_1} = X_{T_2}$). Finally, if $\sum_t y_t \neq 1$, we also have $m(y, X; b) = 0$. These last points, combined with (1.6.12)-(1.6.14), imply $f(b) = 0$. Thus, $b \in B$ and the result follows.

Part 2

The proof is in two steps. First, we show that for all $b \in R$,

$$\operatorname{sgn}(D_1(X;b)) = -\operatorname{sgn}(D_2(X;b)) \quad \text{a.s.} \tag{1.6.15}$$

Second, we show that whenever (1.6.15) holds, we can construct a distribution of $\gamma|X$ such that (1.2.4) holds. The result then follows.

First step: (1.6.15) **holds.**

First, the result holds for $b = \beta_0$ since then $D_j(X;b) = 0$ for $j \in \{1,2\}$. Otherwise, fix $b = c\beta_0 \in R$ and let $\tilde{\lambda} := (1, c, c\lambda_2)$ and $\check{\lambda} := (\lambda_2, c, c\lambda_2)$. Let p (resp. p') denote the minimal number of pairwise coordinate permutations needed to sort the vector $\tilde{\lambda}$ (resp. $\check{\lambda}$) and let $\tilde{\lambda}^s$ (resp. $\check{\lambda}^s$) be the corresponding vector, sorted in ascending order. If $c \in (1/\lambda_2, 1)$, we have $p = 1$ and $p' = 2$, whereas if $c \in (1, \lambda_2)$, $p = 0$ and $p' = 1$. Hence, in all cases, $p' = p + 1$.

Now, for any $x \in \operatorname{Supp}(X)$, notice that

$$D_1(x; b, \lambda) = D_1(x; \beta_0, \tilde{\lambda}) = (-1)^p D_1(x; \beta_0, \tilde{\lambda}^s), \tag{1.6.16}$$

$$D_2(x; b, \lambda) = D_2(x; \beta_0, \check{\lambda}) = (-1)^{p'} D_2(x; \beta_0, \check{\lambda}^s). \tag{1.6.17}$$

Let p'' be the minimal number of pairwise coordinates permutations needed to sort the vector $(x_1'\beta_0, x_2'\beta_0, x_3'\beta_0)$ in ascending order, and let x^s denote the corresponding vector, i.e., such that $x_{s1}'\beta_0 \leq x_{s2}'\beta_0 \leq x_{s3}'\beta_0$. Then

$$D_1(x; \beta_0, \tilde{\lambda}^s) = (-1)^{p''} D_1(x^s; \beta_0, \tilde{\lambda}^s), \tag{1.6.18}$$

$$D_2(x; \beta_0, \check{\lambda}^s) = (-1)^{p''} D_2(x^s; \beta_0, \check{\lambda}^s). \tag{1.6.19}$$

Now, by properties of Chebyshev systems, $D_1(x^s; \beta_0, \tilde{\lambda}^s)$ and $D_2(x^s; \beta_0, \check{\lambda}^s)$ are both non-negative. Moreover, both are nonzero if and only if $|\{x_1'\beta_0, x_2'\beta_0, x_3'\beta_0\}| = 3$. The result follows by (1.6.16)-(1.6.19) and $(-1)^p = -(-1)^{p'}$.

Second step: if (1.6.15) **holds, there exists a distribution of** $\gamma|X$ **such that** (1.2.4) **holds.**

Let us define

$$a_i(\gamma, x) = \frac{w_i \exp(\lambda_i \gamma)}{\prod_{t=1}^{T} \left(1 + \sum_{j=1}^{T-1} w_j \exp(\lambda_j (x_t' \beta_0 + \gamma))\right)}. \tag{1.6.20}$$

Then, we have

$$\mathbb{E}[m(Y, X, b)|X = x] = \mathbb{E}\left[a_1(\gamma, x)|X = x\right] D_1(x, b) + \mathbb{E}\left[a_2(\gamma, x)|X = x\right] D_2(x, b). \tag{1.6.21}$$

Hence, if $D_1(x, b) = D_2(x, b) = 0$, any distribution of $\gamma|X = x$ satisfies $\mathbb{E}[m(Y, X, b)|X = x] = 0$. Now, suppose that $\operatorname{sgn}(D_1(x, b)) = -\operatorname{sgn}(D_2(x, b)) \neq 0$. Then $R(x) := -D_1(x, b)/D_2(x, b) > 0$. Let us define

$$\gamma_0 := \frac{\ln\left[w_1 R(x)/w_2\right]}{\lambda_2 - 1}.$$

Consider for $\gamma|X = x$ the Dirac distribution at γ_0. Then, from (1.6.21), we obtain that $\mathbb{E}[m(Y, X, b)|X = x] = 0$. The result follows.

1.6.6 Theorem 1.3.1

Let us first summarize the proof. We link the current model with a "complete" model where γ is also observed. This model is fully parametric and thus can be analyzed easily. Specifically, we show in a first step that this complete model is differentiable in quadratic mean (see, e.g. van der Vaart, 2000, pp.64-65 for a definition) and has a nonsingular information matrix. In a second step, we establish an abstract expression for the semiparametric efficiency bound. This expression involves in particular the kernel $\mathcal{K}$ of the conditional expectation operator $g \mapsto \mathbb{E}[g(X, Y)|X, \gamma]$. In a third step, we show that

$$\mathcal{K} = \{(x, y) \mapsto q(x) m(x, y; \beta_0), \mathbb{E}[q^2(X)] < \infty\}. \tag{1.6.22}$$

The fourth step of the proof concludes.

First step: the complete model is differentiable in quadratic mean and has a nonsingular information matrix. Let $p(y|x, g; \beta) := \mathbb{P}(Y = y|X = x, \gamma = g; \beta)$. We check that the conditions of Lemma 7.6 in van der Vaart (2000) hold. Under, Assumptions 1.2.1-1.2.2, we have

$$p(y|x, g; \beta) = \prod_{t:y_t=1} F(x_t'\beta + g) \prod_{t:y_t=0} (1 - F(x_t'\beta + g)),$$

where F is C^∞ on $\mathbb{R}$ and takes values in $(0, 1)$. This implies that $\beta \mapsto \ln p(y|x, g; \beta)$ is differentiable. Let $S_\beta := \partial \ln p(Y|X, \gamma; \beta)/\partial\beta$ and let $S_{\beta k}$ denote its k-th component.

We prove that $\mathbb{E}[S_{\beta k}^2] < \infty$. First, remark that

$$S_{\beta k} = \sum_{t=1}^{T} \frac{X_{k,t} F'(X_t'\beta + \gamma)}{[F(X_t'\beta + \gamma)][1 - F(X_t'\beta + \gamma)]} \left[Y_t - F(X_t'\beta + \gamma)\right].$$

Next, we have

$$\begin{aligned} |S_{\beta k}| &\leq \sum_{t=1}^{T} |X_{k,t}| \frac{F'(X_t'\beta + \gamma)}{F(X_t'\beta + \gamma)(1 - F(X_t'\beta + \gamma))} \\ &= \sum_{t=1}^{T} |X_{k,t}| \frac{\sum_{j=1}^{T-1} w_j \lambda_j e^{\lambda_j (X_t'\beta + \gamma)}}{\sum_{j=1}^{T-1} w_j e^{\lambda_j (X_t'\beta + \gamma)}} \\ &\leq \lambda_{T-1} \sum_{t=1}^{T} |X_{k,t}|, \end{aligned} \tag{1.6.23}$$

where we have used the triangle inequality and $|Y_t - F(X_t'\beta + \gamma)| \leq 1$ to obtain the first inequality. Equation (1.6.23) and Assumption 1.2.1.2 imply that $\mathbb{E}[S_{\beta k}^2] < \infty$. By the dominated convergence theorem and again (1.6.23), $\beta \mapsto \mathbb{E}[S_\beta S_\beta']$ is continuous. Therefore, the conditions in Lemma 7.6 in van der Vaart (2000) hold, and the complete model is differentiable in quadratic mean. Moreover,

$$\mathbb{E}[S_\beta S_\beta'] = \mathbb{E}[\mathbb{V}(S_\beta | X, \gamma)] = \sum_{t=1}^{T} \mathbb{E}\left[\left(\frac{F'(X_t'\beta + \gamma)}{[F(X_t'\beta + \gamma)][1 - F(X_t'\beta + \gamma)]} \right)^2 X_t X_t' \right].$$

Then, if for some $v \in \mathbb{R}^K$, $v'\mathbb{E}[S_\beta S_\beta']v = 0$, we would have $X_t'v = 0$ almost surely for all $t \in \{1, \ldots, T\}$. By Assumption 1.3.1.1, this implies $v = 0$. Hence, the information matrix $\mathbb{E}[S_\beta S_\beta']$ is nonsingular.

Second step: $V^\star$ depends on the orthogonal projection of $\mathbb{E}[S_{\beta_0}|X, Y]$ on $\mathcal{K}$. Let $\widetilde{\psi} = (\widetilde{\psi}_1, \ldots, \widetilde{\psi}_K)'$ denote the efficient influence function, as defined p.363 of van der Vaart (2000). Then $V^\star = \mathbb{E}[\widetilde{\psi}\widetilde{\psi}']$ and $\mathbb{E}[\widetilde{\psi}] = 0$. Let $\mathscr{S} :=$span(S_{β_0}), $\mathscr{G} := \{q : \mathbb{E}[q^2(X, \gamma)] < \infty, \mathbb{E}[q(X, \gamma)] = 0\}$ and for any closed convex set A and any $h = (h_1, \ldots, h_K)'$, let Π_A denote the orthogonal projection on A and $\Pi_A(h) = (\Pi_A(h_1), \ldots, \Pi_A(h_K))'$. By Equation (25.29), Lemma 25.34 (since the complete model is differentiable in quadratic mean by the first step) and the same reasoning as in Example 25.36 of van der Vaart (2000), $\widetilde{\psi}$ is the function of (X, Y) of minimal L^2-norm satisfying

$$\widetilde{\chi} = \Pi_{\mathscr{S}+\mathscr{G}}(\widetilde{\psi}), \tag{1.6.24}$$

where $\widetilde{\chi}$ is the efficient influence function of the large model. Because this large model is parametric, we have

$$\widetilde{\chi} = \mathbb{E}[S_{\beta_0} S_{\beta_0}']^{-1} S_{\beta_0}. \tag{1.6.25}$$

Equation (1.6.24) implies $\mathbb{E}[(\widetilde{\psi} - \widetilde{\chi})\widetilde{\chi}'] = 0$. Thus, defining $\ell_{\beta_0} = \mathbb{E}[S_{\beta_0}|Y, X]$, we get

$$\mathbb{E}[\widetilde{\psi}\ell_{\beta_0}'] = \mathbb{E}[\widetilde{\psi} S_{\beta_0}'] = \mathrm{Id}, \tag{1.6.26}$$

Moreover, because $\mathbb{E}[S_{\beta_0}|X,\gamma] = 0$, $\mathscr{S}$ and $\mathscr{G}$ are orthogonal. Thus, (1.6.24) is equivalent to $\Pi_{\mathscr{S}}(\widetilde{\chi}) = \Pi_{\mathscr{S}}(\widetilde{\psi})$ and $\Pi_{\mathscr{G}}(\widetilde{\chi}) = \Pi_{\mathscr{G}}(\widetilde{\psi})$. Moreover, (1.6.25) implies that $\Pi_{\mathscr{G}}(\widetilde{\chi}) = 0$. Hence, $\widetilde{\psi} \in \mathcal{K}^K$. Now, because $\Pi_{\mathcal{K}}$ is an orthogonal projector, we have

$$\mathbb{E}[\widetilde{\psi}\Pi_{\mathcal{K}}(\ell_{\beta_0})'] = \mathbb{E}[\Pi_{\mathcal{K}}(\widetilde{\psi})\ell'_{\beta_0}] = \mathbb{E}[\widetilde{\psi}\ell'_{\beta_0}] = \mathrm{Id},$$

where the last equality follows by (1.6.26). Hence, if $\Pi_{\mathcal{K}}(\ell_{\beta_0})'\lambda = 0$ a.s., we would have $\lambda = 0$. In other words, $\mathbb{E}[\Pi_{\mathcal{K}}(\ell_{\beta_0})\Pi_{\mathcal{K}}(\ell_{\beta_0})']$ is nonsingular. Now, consider the set

$$\mathcal{F} := \left\{\mathbb{E}[\Pi_{\mathcal{K}}(\ell_{\beta_0})\Pi_{\mathcal{K}}(\ell_{\beta_0})']^{-1}\Pi_{\mathcal{K}}(\ell_{\beta_0}) + v :\ \mathbb{E}[v\Pi_{\mathcal{K}}(\ell_{\beta_0})'] = 0\right\}.$$

$\mathcal{F}$ is thus the set of vector-valued functions ψ satisfying the equation $\mathbb{E}[\psi\Pi_{\mathcal{K}}(\ell_{\beta_0})] =$Id. Hence, $\widetilde{\psi}$ being the element of $\mathcal{F}$ with minimum L^2-norm, we obtain

$$\widetilde{\psi} = \mathbb{E}[\Pi_{\mathcal{K}}(\ell_{\beta_0})\Pi_{\mathcal{K}}(\ell_{\beta_0})']^{-1}\Pi_{\mathcal{K}}(\ell_{\beta_0}).$$

Finally, because $V^\star = \mathbb{E}[\widetilde{\psi}\widetilde{\psi}']$,

$$V^\star = \mathbb{E}[\Pi_{\mathcal{K}}(\ell_{\beta_0})\Pi_{\mathcal{K}}(\ell_{\beta_0})']^{-1}. \tag{1.6.27}$$

Third step: (1.6.22) holds. Let $r \in \mathcal{K}$ and let us prove that $r(y,x) = q(x)m(y,x;\beta_0)$ for some q. First, by definition of $\mathcal{K}$, we have, for almost all $(g,x) \in \mathrm{Supp}(\gamma, X)$,

$$\begin{aligned}
0 =& r((0,0,0),x_0) + r((1,0,0),x_0)G(x_1'\beta_0+g) + r((0,1,0),x_0)G(x_2'\beta_0+g)\\
&+r((0,0,1),x_0)G(x_3'\beta_0+g) + r((1,1,0),x_0)G(x_1'\beta_0+g)G(x_2'\beta_0+g)\\
&+r((1,0,1),x_0)G(x_1'\beta_0+g)G(x_3'\beta_0+g) + r((0,1,1),x_0)G(x_2'\beta_0+g)G(x_3'\beta_0+g)\\
&+r((1,1,1),x_0)G(x_1'\beta_0+g)G(x_2'\beta_0+g)G(x_3'\beta_0+g).
\end{aligned} \tag{1.6.28}$$

Let $a_t := x_t'\beta_0$ for $t \in \{1,2,3\}$ and, for the sake of conciseness, let us remove the dependence of r on x. Then, using Assumption 1.2.2, we obtain, for almost all (g,x),

$$\begin{aligned}
0 =& A_1e^{0\times g} + A_2e^{g} + A_3e^{\lambda_2 g} + A_4e^{2g} + A_5e^{2\lambda_2 g} + A_6e^{(1+\lambda_2)g} + A_7e^{3g} + A_8e^{(2+\lambda_2)g}\\
&+ A_9e^{(1+2\lambda_2)g} + A_{10}e^{3\lambda_2 g},
\end{aligned}$$

where

$$
\begin{aligned}
A_1 :=& r(0,0,0),\\
A_2 :=& w_1\left[r(1,0,0)e^{a_1}+r(0,1,0)e^{a_2}+r(0,0,1)e^{a_3}\right],\\
A_3 :=& w_2\left[r(1,0,0)e^{\lambda_2 a_1}+r(0,1,0)e^{\lambda_2 a_2}+r(0,0,1)e^{\lambda_2 a_3}\right],\\
A_4 :=& w_1^2\left[r(1,1,0)e^{(a_1+a_2)}+r(1,0,1)e^{(a_1+a_3)}+r(0,1,1)e^{(a_2+a_3)}\right],\\
A_5 :=& w_1w_2\left[r(1,1,0)(e^{a_1+\lambda_2 a_2}+e^{a_2+\lambda_2 a_1})+r(1,0,1)(e^{a_1+\lambda_2 a_3}+e^{a_3+\lambda_2 a_1})\right.\\
&\left.+r(0,1,1)(e^{a_2+\lambda_2 a_3}+e^{a_3+\lambda_2 a_2})\right],\\
A_6 :=& w_2^2\left[r(1,1,0)e^{\lambda_2(a_1+a_2)}+r(1,0,1)e^{\lambda_2(a_1+a_3)}+r(0,1,1)e^{\lambda_2(a_2+a_3)}\right],\\
A_7 :=& w_1^3 r(1,1,1)e^{a_1+a_2+a_3},\\
A_8 :=& w_1^2w_2 r(1,1,1)\left[e^{a_1+a_2+\lambda_2 a_3}+e^{a_1+\lambda_2 a_2+a_3}+e^{\lambda_2 a_1+a_2+a_3}\right],\\
A_9 :=& w_1w_2^2 r(1,1,1)\left[e^{a_1+\lambda_2(a_2+a_3)}+e^{a_2+\lambda_2(a_1+a_3)}+e^{a_3+\lambda_2(a_1+a_2)}\right],\\
A_{10} :=& w_2^3 r(1,1,1)e^{\lambda_2(a_1+a_2+a_3)}.
\end{aligned}
$$

Since $\lambda_2 = 2$ is excluded by assumption, there are three cases left depending on the number of different exponents in Equation (1.6.28).

First, we consider $\lambda_2 \notin \{3/2, 3\}$. By Lemma 1.6.1 and because $|\mathrm{Supp}(\gamma|X)| \geq 10$, we obtain $A_k = 0$ for all $k \in \{1, \ldots, 10\}$. $A_1 = A_7 = 0$ imply that $r(0,0,0) = r(1,1,1) = 0$. Next, $A_4 = A_6 = 0$ implies that either $r(1,0,1) = r(1,1,0) = r(0,1,1) = 0$ or

$$
\begin{cases}
r(1,1,0) &= -r(1,0,1)e^{\lambda_2(a_3-a_2)} - r(0,1,1)e^{\lambda_2(a_3-a_1)},\\
r(1,1,0) &= -r(1,0,1)e^{(a_3-a_2)} - r(0,1,1)e^{(a_3-a_1)}.
\end{cases}
\tag{1.6.29}
$$

Consider the second case. $A_5 = 0$ implies, since $(r(1,0,1), r(1,1,0), r(0,1,1)) \neq (0,0,0)$,

$$
r(1,1,0) = -r(1,0,1)\frac{e^{a_1+\lambda_2 a_3}+e^{a_3+\lambda_2 a_1}}{e^{a_1+\lambda_2 a_2}+e^{a_2+\lambda_2 a_1}} - r(0,1,1)\frac{e^{a_2+\lambda_2 a_3}+e^{a_3+\lambda_2 a_2}}{e^{a_1+\lambda_2 a_2}+e^{a_2+\lambda_2 a_1}}.
$$

By assumption, for almost every $x = (x_1, x_2, x_3)$, $a_3 \neq a_2$ and $a_3 \neq a_1$. Then, using the latter display with equation (1.6.29) yields, since $\lambda_2 \neq 1$,

$$
\begin{aligned}
r(1,0,1) &= r(0,1,1)\left[e^{\lambda_2(a_3-a_2)} - e^{a_3-a_2}\right]^{-1}\left[e^{a_3-a_1} - e^{\lambda_2(a_3-a_1)}\right],\\
r(1,0,1) &= r(0,1,1)\left[e^{\lambda_2(a_3-a_2)} - \frac{e^{a_1+\lambda_2 a_3}+e^{a_3+\lambda_2 a_1}}{e^{a_1+\lambda_2 a_2}+e^{a_2+\lambda_2 a_1}}\right]^{-1}\\
&\quad\times\left[\frac{e^{a_2+\lambda_2 a_3}+e^{a_3+\lambda_2 a_2}}{e^{a_1+\lambda_2 a_2}+e^{a_2+\lambda_2 a_1}} - e^{\lambda_2(a_3-a_1)}\right].
\end{aligned}
$$

Since $(r(1,1,0), r(1,0,1), r(0,1,1)) \neq (0,0,0)$, these equalities and (1.6.29) imply that $r(1,0,1) \neq 0$ and $r(0,1,1) \neq 0$. Then

$$\frac{e^{(1-\lambda_2)a_2}}{e^{(1-\lambda_2)a_1}} \frac{e^{a_3+\lambda_2 a_2+(\lambda_2-1)a_1} - e^{\lambda_2(a_2+a_3)}}{e^{\lambda_2(a_1+a_2)} - e^{(\lambda_2-1)a_2+\lambda_2 a_1+a_3}} = \frac{e^{a_3+\lambda_2 a_2+(\lambda_2-1)a_1} - e^{\lambda_2(a_2+a_3)}}{e^{\lambda_2(a_1+a_2)} - e^{(\lambda_2-1)a_2+\lambda_2 a_1+a_3}},$$

which is equivalent to $a_1 = a_2$. By assumption, the set of x for which this occurs is of probability zero. In other words, for almost every x,

$$r((1,1,0),x) = r((1,0,1),x) = r((0,1,1),x) = 0.$$

$A_2 = A_3 = 0$ implies that either $r(1,0,0) = r(0,1,0) = r(0,0,1) = 0$ or

$$\begin{cases} r(0,0,1) &= -e^{(a_1-a_3)} r(1,0,0) - e^{(a_2-a_3)} r(0,1,0), \\ r(0,0,1) &= -e^{\lambda_2(a_1-a_3)} r(1,0,0) - e^{\lambda_2(a_2-a_3)} r(0,1,0). \end{cases}$$

In the first case, almost surely $r(Y,X) = 0 = 0 \times m(Y,X;\beta_0)$. In the second case, $r(Y,X) = q(X) \times m(Y,X;\beta_0)$ for some $q \in L^2_X$. The result follows.

Now, we turn to $\lambda_2 = 3/2$. Then, for almost all $(g,x) \in \text{Supp}(\gamma, X)$,

$$0 = A_1 e^{0\times g} + A_2 e^{g} + A_3 e^{\frac{3}{2}g} + A_4 e^{2g} + (A_5 + A_7) e^{3g} + A_6 e^{\frac{5}{2}g} + A_8 e^{\frac{7}{2}g} + A_9 e^{4g} + A_{10} e^{\frac{9}{2}g}.$$

By Lemma 1.6.1 and because $|\text{Supp}(\gamma|X)| \geq 9$, we obtain $A_5 + A_7 = 0$ and $A_k = 0$ for all $k \notin \{5,7\}$. $A_1 = A_{10} = 0$ implies that $r(0,0,0) = r(1,1,1) = 0$ which in turn implies that $A_7 = 0$ and thus $A_5 = 0$. Hence, we have $A_k = 0$ for all $k \in \{1,\ldots,10\}$ and the same reasoning as when $\lambda_2 \notin \{3/2, 3\}$ allows us to obtain the result.

Finally, we consider $\lambda_2 = 3$. Then, for all (g,x),

$$0 = A_1 e^{0\times g} + A_2 e^{g} + (A_3 + A_7) e^{3g} + A_4 e^{2g} + A_5 e^{6g} + A_6 e^{4g} + A_7 e^{5g} + A_8 e^{7g} + A_9 e^{9g},$$

By Lemma 1.6.1 and because $|\text{Supp}(\gamma|X)| \geq 9$, we obtain $A_3 + A_7 = 0$ and $A_k = 0$ for all $k \notin \{3,7\}$. $A_1 = A_{10} = 0$ implies that $r(0,0,0) = r(1,1,1) = 0$ which in turn implies that $A_7 = 0$ and thus $A_3 = 0$. Hence, $A_k = 0$ for all $k \in \{1,\ldots,10\}$ and the result follows again as when $\lambda_2 \notin \{3/2, 3\}$.

Fourth step: conclusion. By Steps 2 and 3, there exists $q_0(X)$ such that $\Pi_{\mathcal{K}}(\ell_{\beta_0}) = q_0(X) m(Y,X;\beta_0)$. Moreover, by definition of the orthogonal projection, $\Pi_{\mathcal{K}}(\ell_{\beta_0}) - \ell_{\beta_0} \in (\mathcal{K}^{\perp})^K$. Hence, again by Step 3, we have, for all $q \in L^2_X$,

$$\mathbb{E}[q_0(X) q(X) m(Y,X;\beta_0)^2] = \mathbb{E}[\ell_{\beta_0} q(X) m(Y,X;\beta_0)].$$

This implies that

$$q_0(X)\Omega(X) = \mathbb{E}[\ell_{\beta_0} m(Y,X;\beta_0)|X].$$

As a result, because $\ell_{\beta_0} = \mathbb{E}[S_{\beta_0}|Y,X]$,

$$\begin{aligned}\Pi_{\mathcal{K}}(\ell_{\beta_0}) =& \Omega^{-1}(X)m(Y,X;\beta_0)\mathbb{E}[\ell_{\beta_0}m(Y,X;\beta_0)|X] \\ =& \Omega^{-1}(X)m(Y,X;\beta_0)\mathbb{E}[S_{\beta_0}m(Y,X;\beta_0)|X].\end{aligned}$$

Then, using (1.6.27), we obtain

$$V^\star = \mathbb{E}\left[\Omega^{-1}(X)\mathbb{E}[S_{\beta_0}m(Y,X;\beta_0)|X]\mathbb{E}[S_{\beta_0}m(Y,X;\beta_0)|X]'\right]^{-1}.$$

Now, by the end of the proof of Theorem 1.2.3, we have, for all β,

$$0 = \mathbb{E}_\beta\left[m(Y,X;\beta)|X,\gamma\right].$$

As a result,

$$\begin{aligned}0 &= \nabla_\beta \mathbb{E}_\beta\left[m(Y,X;\beta)|X,\gamma\right] \\ &= \mathbb{E}_\beta\left[\nabla_\beta m(Y,X;\beta)|X,\gamma\right] + \mathbb{E}_\beta\left[m(Y,X;\beta)S_\beta|X,\gamma\right].\end{aligned}$$

Evaluating this equality at β_0 and integrating over γ yields:

$$\mathbb{E}[S_{\beta_0}m(Y,X;\beta_0)|X] = -\mathbb{E}[\nabla_\beta m(Y,X;\beta_0)|X] = -R(X).$$

We conclude that

$$V^\star = \mathbb{E}\left[\Omega^{-1}(X)R(X)R(X)'\right]^{-1} = V_0,$$

which is a well-defined matrix by Assumption 1.3.1.1.

1.7 Extensions

We show that there exists a class of distribution functions even larger, for which non-trivial moment conditions exist and allow for identification and $\sqrt{n}$-consistent estimation of β_0. To the best of our knowledge, this class is new to the existing literature. Let Λ_τ denote a subset of $\{(\lambda_1,\ldots,\lambda_\tau)\in\mathbb{R}^\tau : \prod_{r\neq s}(\lambda_r-\lambda_s)\neq 0\}$.

Assumption 1.7.1 (Exponential polynomial odds-ratio distributions) *There exist known* $(\tau_1,\tau_2)\in\mathbb{N}^{*2}$, $w=(w_1,\ldots,w_{\tau_1})\in\mathbb{R}^{\tau_1}$, $w'=(w'_1,\ldots,w'_{\tau_2})\in\mathbb{R}^{\tau_2}$, *and* $\lambda=(\lambda_1,\ldots,\lambda_{\tau_1\vee\tau_2})\in\Lambda_{\tau_1\vee\tau_2}$ *such that*

$$\frac{F(\varepsilon)}{1-F(\varepsilon)} = \frac{\sum_{j=1}^{\tau_1} w_j\exp(\lambda_j\varepsilon)}{\sum_{j=1}^{\tau_2} w'_j\exp(-\lambda_j\varepsilon)}. \tag{1.7.1}$$

defines a cumulative distribution function F *on* $\mathbb{R}$.

Assumption 1.7.1 generalizes both types of generalized logit distributions proposed in Assumption 1.2.2. The first type is obtained by letting $\tau_2=\tau_1+1$, $w'_1=\cdots=w'_{\tau_1}=0$, $w'_{\tau_1+1}=1$, $\lambda_{\tau_1+1}=0$, and $\lambda_1=1$; the second type, by letting $\tau_1=\tau_2+1$,

$w_1 = \cdots = w_{\tau_2} = 0$, $w_{\tau_2+1} = 1$, $\lambda_{\tau_2+1} = 0$, and $\lambda_1 = 1$. Assumption 1.7.1 also allows for any mixture of logit distributions. In particular, beyond the standard logistic distribution, it covers the mixture of logit distributions considered in Honoré and Weidner (2020)'s equation (12) when $\tau_1 = \tau_2 = 2$ and

$$\begin{aligned}(w_1, w_2) &= (1, \omega \exp(\mu_2) + (1-\omega)\exp(\mu_1)),\\ (\lambda_1, \lambda_2) &= (\lambda, 0),\\ (w_1', w_2') &= (\exp(\mu_1+\mu_2), (1-\omega)\exp(\mu_2) + \omega\exp(\mu_1)).\end{aligned}$$

Theorem 1.7.1 *Let Assumptions 1.2.1 and 1.7.1 hold. Then, for T sufficiently large, there exists a non-trivial function m of $(Y, X, \beta, \tau_1, \tau_2, w, w', \lambda)$ such that*

$$\mathbb{E}[m(Y, X; \beta_0)|X, \gamma] = \mathbb{E}[m(Y, X; \beta_0)|X] = 0. \tag{1.7.2}$$

Proof: Let $G := F/(1-F)$, $z_t := x_t'\beta_0$, $\pi_t := \exp(z_t)$, and $a := \exp(\gamma)$. We have

$$\begin{aligned}p(y|x,\gamma) &:= \prod_{t=1}^T [1-F(z_t+\gamma)]^{1-y_t}[F(z_t+\gamma)]^{y_t}\\ &= \prod_{t=1}^T \left(\frac{1}{1+G(z_t+\gamma)}\right)^{1-y_t}\left(\frac{G(z_t+\gamma)}{1+G(z_t+\gamma)}\right)^{y_t}\\ &= \left(\prod_{t=1}^T \frac{1}{1+G(z_t+\gamma)}\right)\left(\prod_{s:y_s=1} G(z_s+\gamma)\right)\\ &= \left(\prod_{t=1}^T \frac{1}{1+G(\log(\pi_t a))}\right)\left(\prod_{s:y_s=1} \frac{\sum_{j=1}^{\tau_1} w_j(\pi_s a)^{\lambda_j}}{\sum_{j=1}^{\tau_2} w_j'(\pi_s a)^{-\lambda_j}}\right)\\ &= \left(\prod_{t=1}^T \frac{1}{1+G(\log(\pi_t a))}\right)\left(\prod_{s=1}^T \frac{1}{\sum_{j=1}^{\tau_2} w_j'(\pi_s a)^{-\lambda_j}}\right)\\ &\quad \times \left(\prod_{s:y_s=1}\sum_{j=1}^{\tau_1} w_j(\pi_s a)^{\lambda_j}\right)\left(\prod_{r:y_r=0}\sum_{j=1}^{\tau_2} w_j'(\pi_r a)^{-\lambda_j}\right)\\ &=: \kappa(a)p(y,a),\end{aligned}$$

where

$$\begin{aligned}\kappa(a) &= \left(\prod_{t=1}^T \frac{1}{1+G(\log(\pi_t a))}\right)\left(\prod_{s=1}^T \frac{1}{\sum_{j=1}^{\tau_2} w_j'(\pi_s a)^{-\lambda_j}}\right),\\ p(y,a) &= \left(\prod_{s:y_s=1}\sum_{j=1}^{\tau_1} w_j(\pi_s a)^{\lambda_j}\right)\left(\prod_{r:y_r=0}\sum_{j=1}^{\tau_2} w_j'(\pi_r a)^{-\lambda_j}\right).\end{aligned}$$

Let $\mathcal{J}_1 = \{j : w_j \neq 0\}$ and $\mathcal{J}_2 = \{j : w'_j \neq 0\}$. Thus, $p(y, a)$ is an exponential polynomial in γ with at most

$$K := \left| \left\{ \sum_{j \in \mathcal{J}_1} A_j \lambda_j - \sum_{j \in \mathcal{J}_2} B_j \lambda_j : (A_j, B_j) \in \{0, \dots, T\}^2, \sum_{j \in \mathcal{J}_1} A_j + \sum_{j \in \mathcal{J}_2} B_j = T \right\} \right|,$$

distinct coefficients, i.e., there exist K distinct real numbers $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_K$ and $(c_1(y), \dots, c_K(y)) \in \mathbb{R}^K$ such that

$$p(y, a) = \sum_{k=1}^{K} c_k(y) \exp(\widetilde{\lambda}_k \gamma).$$

We are looking for non-trivial moment functions $m(y) \in \mathbb{R}$ such that

$$\sum_{y \in \{0,1\}^T} m(y) \kappa(a) p(y, a) = 0, \quad \forall a \in (0, +\infty).$$

Because $\kappa(a) > 0$, this is equivalent to finding $m(y) \in \mathbb{R}$ that satisfy

$$\sum_{k=1}^{K} \sum_{y \in \{0,1\}^T} m(y) c_k(y) \exp(\widetilde{\lambda}_k \gamma) = 0, \quad \forall \gamma \in \mathbb{R}.$$

By Lemma 1.6.1, this is equivalent to finding $m(y) \in \mathbb{R}$ that satisfy

$$\sum_{y \in \{0,1\}^T} m(y) c_k(y) = 0, \quad \forall k \in \{1, \dots, K\}. \tag{1.7.3}$$

These are K linear conditions in 2^T unknown parameters $m(y)$. We therefore have at least $2^T - K$ linear independent solutions $m(y)$. Let us show that $2^T - K > 0$ for T sufficiently large. Let $\tau_1^* = |\mathcal{J}_1|$ and $\tau_2^* = |\mathcal{J}_2|$. It is easy to see that K is bounded above by

$$\sum_{s=0}^{T} \binom{s + \tau_1^* - 1}{s} \binom{T - s + \tau_2^* - 1}{T - s}.$$

For any $(a, b) \in \mathbb{N}^2$ such that $a + b \geq 1$, we have

$$\binom{a + b - 1}{a} \leq \frac{(a + b - 1)^{b-1}}{(b - 1)!}.$$

Hence, for any $0 \leq s \leq T$,

$$\begin{aligned} \binom{s + \tau_1^* - 1}{s} \binom{T - s + \tau_2^* - 1}{T - s} &\leq \frac{(s + \tau_1^* - 1)^{\tau_1^* - 1} (T - s + \tau_2^* - 1)^{\tau_2^* - 1}}{(\tau_1^* - 1)! (\tau_2^* - 1)!} \\ &\leq (T + \tau_1^* + \tau_2^* - 1)^{\tau_1^* + \tau_2^* - 2}. \end{aligned}$$

For T sufficiently large, $2^T - K \geq 2^T - (T+1)(T + \tau_1^* + \tau_2^* - 1)^{\tau_1^* + \tau_2^* - 2} > 0$ and there exists a non-trivial moment function m that verifies (1.7.2). □

Chapter 2

Identification and (Fast) Estimation of Large Nonlinear Panel Models with Two-Way Fixed Effects

Il fallut que Colomb partît avec des fous pour découvrir l'Amérique. Et voyez comme cette folie a pris corps, et duré.

André Breton, *Manifeste du surréalisme*

Abstract: We study a nonlinear two-way fixed effects panel model that allows for unobserved individual heterogeneity in slopes and flexibly specified link function. The former is relevant when the researcher is interested in the distributional causal effects of covariates, and the latter mitigates potential misspecification errors due to restrictions imposed on the link function. We show that the fixed effects parameters and the link function can be identified when both individual and time dimensions are large. We propose a novel iterative Gauss-Seidel estimation procedure that overcomes the practical challenge of dimensionality in the number of fixed effects when the dataset is large. We revisit two empirical studies in trade (Helpman et al., 2008) and innovation (Aghion et al., 2013), and find non-negligible unobserved dispersion in trade elasticity across countries and the effect of institutional ownership on innovation across firms. These exercises emphasize the usefulness of our method in capturing flexible heterogeneity in the causal relationship of interest that may have important implications for the subsequent policy analysis.[1]

[1]This chapter is based on a co-authored paper with Ao Wang (University of Warwick).

2.1 Introduction

Nonlinear two-way fixed effects panel models are gaining popularity in economic research. This class of models typically features individual and time dimensions, enabling researchers to incorporate rich heterogeneity in empirical research.[2] Technically, by allowing the two dimensions to increase to infinity, one can reduce the incidental parameter problem in panel data models (Lancaster, 2000; Neyman and Scott, 1948) to a post-estimation bias correction (Fernández-Val and Weidner, 2016). The application of these models, however, is still subject to theoretical and practical challenges. First, the extent to which the model is nonparametrically identified is unclear, leaving many parametric assumptions in empirical research, such as common slope parameters across individuals/time and parametrically specified error terms, unjustified. Second, even in a parametric setting, the routine estimation procedure (e.g., concentrated MLE) is subject to a challenge of dimensionality: the number of fixed effects can be too large to be handled in reasonable time.[3] This dimensionality is particularly difficult to deal with when the dataset is large and/or the researcher wishes to incorporate multiple dimensions of unobserved heterogeneity.

In this chapter, we tackle these two challenges in a class of index function static models characterized by the probability of individual $i = 1, \dots, N$ from choosing $y_{it} \in \mathcal{Y}$ at time $t = 1, \dots, T$:

$$\mathbb{P}(y_{it} = y | x_{i1}, \dots, x_{it}, \alpha_i, \beta_i, \xi_t) = g(y; \alpha_i + \xi_t + x_{it}'\beta_i), \tag{2.1.1}$$

where x_{it} are individual i's observed characteristics at time t, (α_i, β_i) are individual-fixed effects, ξ_t is a time-fixed effect, and g is a link function. This model encompasses settings with a single index, such as binary outcome, ordered outcome and count outcome, as well as those with multimodal outcome, and has been widely used in literature including empirical industrial organization (Dubois et al., 2020), international trade (Helpman et al., 2008), labor (Abowd et al., 1999), innovation (Aghion et al., 2013), and network (Jochmans, 2018). Compared to routinely used nonlinear two-way fixed effects models, model (2.1.1) features two important relaxations. First, we allow the slope parameter, β_i, to be individual-specific rather than common across individuals ($\beta_i = \beta$).[4] This feature enables applied researchers to incorporate (unobserved) heterogeneity in the causal effect of covariates of interest across individuals, e.g., household's price sensitivity, trade elasticity. This is particularly relevant when the researcher is interested in the distributional causal effects of covariates and

[2] In some situations, the time dimension also refers to the same set of individuals in the individual dimension. Then, the panel data describes interactions between two individuals, e.g., export and import (Helpman et al., 2008) and directed network (Jochmans, 2018).

[3] For instance, the standard implementation of the MLE requires storing and inverting a Hessian matrix whose size is equal to the number of parameters including the fixed effects. This numerical challenge can be even more severe if one wishes to implement the post-estimation bias correction. See Section 2.3.2 for details.

[4] We can also allow for the slope parameter to be time-specific rather than individual-specific. The main results still hold. See Remark 2.2.1 for details.

their implications for policy evaluations. Second, the link function g can be left to be estimated (e.g., up to a finite number of parameters) rather than specified as a known function (e.g., logit, probit). This flexibility enables to mitigate potential misspecification errors due to restrictions made on the link function.

The central theoretical question in this chapter is then the extent to which the parameters in model (2.1.1) are identified under such relaxations. In the setting of large N and large T, we prove that β_i, α_i, ξ_t, and function g can be point-identified, a novel identification result in the literature of panel data methods. Our strategy relies on the technique of *compensating variable*.[5] Intuitively speaking, we require the existence of a variable in x_{it} that can compensate the variation due to other components of the index (other covariates in x_{it}, α_i, and ξ_t) to keep the index unchanged. Under standard assumptions on the link function (e.g., monotonicity with respect to the single index), one can then back out the amount of the compensation, giving rise to restrictions on the parameters of interest and achieving the point identification. When the support of (β_i, α_i) and/or ξ_t is large, a large support condition on the compensating variable (e.g., $\mathbb{R}$) is needed to apply our strategy. Otherwise, it can still apply with a limited-support compensating variable (e.g., an interval) when the ranges/shape of parameters of interest is limited, an appealing feature to practitioners who may have prior knowledge of (β_i, α_i) and ξ_t. Moreover, it allows for other variables to be endogenous, e.g., correlated with time-fixed effect ξ_t.

The identification result provides a theoretical ground for a semi(or non)-parametric estimation of model (2.1.1), especially when the data are rich in both individual and time dimensions and/or in the presence of multiple dimensions of individual heterogeneity. To deal with the emerging challenge of dimensionality in estimation, we propose a novel iterative Gauss-Seidel procedure to implement the likelihood fixed effects estimators routinely used in the literature. During each iteration, we sequentially update the estimates of individual fixed effects, time fixed effects, and common parameters. Different from the usual Gauss-Seidel procedure, we leverage the structure of the separable two-way fixed effects to update the estimated individual (time) fixed effects in a fully parallelized way across N individuals (T time periods). This substantially alleviates the computational burden of concentrating out a potentially large number of fixed effects in the usual implementation of the MLE. Besides, the proposed procedure has desired theoretical and numerical properties. First, we prove that under standard conditions, the resulting estimators converge to the MLE ones when the number of iterations is large enough. This numerical equivalence legitimizes the use of the proposed procedure in inference.[6] Second, extensive Monte Carlo simulations suggest its fast convergence. It already achieves a good numerical

[5] This term was first introduced by Hicks (1939) and later appears in Lewbel (2019)'s survey of identification in econometrics.

[6] We also provide a consistency result regarding the MLE estimator of β_i that implies the consistency of the plug-in estimators for the distributional features of β_i. See the discussion of inference in Section 2.3.2 and Appendix 2.7.3. Besides, we develop a practical Bootstrap construction of confidence intervals for such features. See Section 2.4 for the Monte Carlo evidence of its good finite-sample performance.

approximation to the MLE estimator after a few iterations using only a fractional execution time of existing methods (e.g., STATA command `logitfe`). Finally, this iterative estimation procedure with parallelized updating is of general interest. It can be applied to both moderate and large-T settings, and conveniently augmented with post-estimation bias reduction and bias correction. One can also extend it to a panel model along the lines of (2.1.1) with multi-way fixed effects.[7] We provide a Python package `nlmfe` that implements this procedure and some of the extensions.[8]

To demonstrate the empirical relevance of the proposed method, we revisit two classic empirical studies in international trade (Helpman et al., 2008) and innovation (Aghion et al., 2013). Specifically, we investigate the extent to which the causal effect of interest is heterogeneous across individuals. In other words, different from the two-way fixed effects models with common slope parameter $\beta_i = \beta$ in the original papers, we allow for individual-specific slopes that encapsulate potentially heterogeneous causal effects of covariates and explore the underlying mechanisms. In the setting of Helpman et al. (2008), we specify country-specific rather than constant trade elasticity; in the setting of Aghion et al. (2013), we allow for a firm's innovation to react differently to the same change in institutional ownership. In both illustrations, we find non-negligible dispersion in the estimated slopes across countries/firms. This dispersion suggests intuitive distributional patterns of the heterogeneity in the causal relationship explained by observed characteristics of countries/firms. The residual dispersion in the slopes (i.e., unexplained by the observed characteristics), however, is still significant. These exercises emphasize the usefulness of the proposed method in capturing flexible (and unobserved) heterogeneity in the causal relationship of interest.

Related literature. This chapter contributes to the literature on nonseparable panel data models with unobserved fixed effects. Recent progresses on two-way fixed effects models with large N and T mostly focus on estimation and inference, while there are much fewer results on identification.[9] To the best of our knowledge, we are the first to provide a systematic treatment of the identification of nonlinear two-way fixed effects index models when both N and T are large, serving as a theoretical foundation for their estimation and inference. Our identification strategy relies on the arguments of compensating variable, a technique that has been used in the literature but different contexts from ours (see D'Haultfoeuille et al. (2021) and D'Haultfœuille et al. (2022)). Several recent papers study models of network formation with two-way fixed effects, and leverage specific features of the setting (e.g., the dimension of

[7] See also Iaria and Wang (2021) for an application of a similar iterative procedure to estimating demand with large choice sets.

[8] The package is available at https://github.com/martinmugnier/nlmfe.

[9] For the progress on estimation and inference, see Hahn and Kuersteiner (2002), Hahn and Newey (2004), Fernandez-Val (2009), Dhaene and Jochmans (2015), Chen (2016), Fernández-Val and Weidner (2016, 2018), Chen et al. (2021) among others.

time completely coincides with that of individuals) to achieve identification.[10] We focus on a different setting in which the relationship between time and individual dimensions is unrestricted.[11] Relative to the literature of classic nonlinear fixed-T panel models, our results articulate that all the structural parameters (fixed effects, slope parameters, link function) can be nonparametrically identified when both N and T are large, which is hard to obtain without further restrictions when T is fixed (Chamberlain, 2010).[12] In addition, our methodology applies not only to the usual setting with common slope parameters (e.g., Fernández-Val and Weidner, 2016) but also the case with heterogeneous slopes across individuals/time. Boneva and Linton (2017) study estimation and inference of a nonlinear panel model with interactive fixed effects and heterogeneous slopes when T increases at a rate slower than N. Differently, our asymptotic results focus on the setting when T and N increase at the same rate.

Our work also contributes to the literature on the estimation of nonlinear fixed effects models. Recent progresses in this literature rely on the (multinomial) logit structure (Charbonneau, 2017; D'Haultfoeuille and Iaria, 2016; Graham, 2017; Stammann et al., 2016), focus on specific models such as Poisson (Correia et al., 2020) and Generalized Linear Models (Hinz et al., 2019), use alternating projection and Frisch-Waugh-Lowell methods (Czarnowske and Stammann, 2020; Gaure, 2013; Stammann, 2018; Stammann et al., 2016), EM method (Chen, 2016), or Minorization-Maximation algorithm (Chen et al., 2021) to alleviate the numerical bottleneck due to many fixed effects in the MLE. Differently, we provide a general Gauss-Seidel estimation procedure to tackle this challenge of dimensionality and improve upon existing Gauss-Seidel approaches in several aspects. Hospido (2012) adopts the Gauss-Seidel algorithm in the estimation of nonlinear models with only individual fixed effects. Guimaraes and Portugal (2010) employ such algorithm to estimate a linear model with many fixed effects and note that their approach can be considerably slowed down when applied to estimating nonlinear models.[13] Different from these approaches, our Gauss-Seidel procedure leverages the separable two-way fixed effects structure, parallelizing the updatings of individual and time fixed effects and significantly reducing the computational complexity of the concentration step in the MLE. Bergé (2018)'s approach sequentially updates all fixed effects and guarantees the likelihood is increasing (but not necessarily to the global maximum) in the concentration step. Instead, we establish the numerical equivalence of our procedure to the fixed effect MLE under standard

[10] See Graham (2017), Toth (2017), Gao (2020), Zeleneev (2020), Candelaria (2020) for examples. See also de Paula (2020) for a review of recent progress.

[11] Jochmans (2018) studies directed network formation and his setting is close to ours. Differently, he focuses on inference on common parameters rather than identification.

[12] Altonji and Matzkin (2005) obtain identification of the structural function in a fixed-T setting with individual-fixed effects using restrictions on the conditional distribution on the fixed effects (see their Assumption 4.4). Having large-T also allows to relax conditions needed to guarantee desired asymptotic properties of the estimator. One such condition is time homogeneity. See Athey and Imbens (2006), Evdokimov (2010, 2011), Hoderlein and White (2012), Chernozhukov et al. (2013), Botosaru and Muris (2017a) for examples using this condition in fixed-T setting.

[13] See page 16 in their paper.

conditions, ensuring its validity in inference and robust finite-sample performance.

Organization. Section 2.2 introduces necessary notations and the model. The main results of identification and estimation are in Section 2.3. Section 2.4 summarizes the results of Monte Carlo simulations. Two empirical illustrations are included in Section 2.5. All proofs and additional results can be found in Appendices 2.7.1-2.7.4.

2.2 Model

We consider a class of index function models with discrete outcome characterized by the probability of individual $i \in \mathbf{N}$ at time $t \in \mathbf{T}$ choosing $y_{it} \in \mathcal{Y}$:

$$\mathbb{P}\left(y_{it} = y | x_{i1}, \ldots, x_{it}, \alpha_i, \beta_i, \xi_t\right) = g(y; \alpha_i + \xi_t + x_{it}'\beta_i), \tag{2.2.1}$$

where x_{it} are individual i's observed characteristics at time t, (α_i, β_i) are individual-fixed effects, ξ_t is a time-fixed effect, and g is a link function of outcome y and index $\alpha_i + \xi_t + x_{it}'\beta_i$. Model (2.2.1) differs from routinely used two-way fixed effects models in two ways. First, it allows for individual-specific slopes β_i, rather than common slope $\beta_i = \beta$, that capture heterogeneous causal effect of covariates x_{it} across individuals and are potentially unobserved to the researcher, such as household's heterogeneous sensitivity to price change.[14] Second, the link function g in model (2.2.1) can be flexibly specified (e.g., unknown up to a finite number of parameters to be estimated), relaxing the common restrictions such as probit and logit in empirical research.

Remark 2.2.1 *One can consider a model with time-specific slope parameters:*

$$\mathbb{P}\left(y_{it} = y | x_{i1}, \ldots, x_{it}, \alpha_i, \beta_t, \xi_t\right) = g(y; \alpha_i + \xi_t + x_{it}'\beta_t). \tag{2.2.2}$$

The parameter β_t captures potentially heterogeneous causal effect of x_{it} across time periods. To simplify the exposition, we will focus on model (2.2.1) *in the main text and extend our main results to model* (2.2.2) *in Section 2.7.8.*

Before proceeding with the identification and estimation, we provide some leading examples.

Example 1 (Binary outcome)

$$y_{it} = \mathbf{1}\{\alpha_i + \xi_t + x_{it}'\beta_i - u_{it} > 0\},$$

[14] One can use an it-specific β_{it} and specify $\beta_{it} = \gamma z_{it}$ to capture observed heterogeneity in slopes, where z_{it} is a vector of observed characteristics of individual i at time t. This is equivalent to adding $x_{it}z_{it}$ in (2.2.1) with common slopes γ across individuals and time periods.

where $(x_{i1}', \ldots, x_{it}', \alpha_i, \beta_i', \xi_t)'$ *and* u_{it} *are independent, and* u_{it} *is distributed according to a cumulative distribution function (cdf)* F*. Then,*

$$g(y; \alpha_i + \xi_t + x_{it}'\beta_i) = \mathbf{1}\{y = 1\}F(\alpha_i + \xi_t + x_{it}'\beta_i) + \mathbf{1}\{y = 0\}(1 - F(\alpha_i + \xi_t + x_{it}'\beta_i)).$$

Example 2 (Ordered outcome)

$$y_{it} = \begin{cases} 0 & \text{if } \alpha_i + \xi_t + x_{it}'\beta_i - u_{it} < d_1. \\ 1 & \text{if } d_1 \leq \alpha_i + \xi_t + x_{it}'\beta_i - u_{it} < d_2. \\ 2 & \text{if } \alpha_i + \xi_t + x_{it}'\beta_i - u_{it} \geq d_2, \end{cases}$$

where $d_2 > d_1$*,* $(x_{i1}', \ldots, x_{it}', \alpha_i, \beta_i', \xi_t)'$ *and* u_{it} *are independent, and* u_{it} *is distributed according to a cdf* F*. Then,*

$$g(y; \alpha_i + \xi_t + x_{it}'\beta_i) = \begin{cases} 1 - F(\alpha_i + \xi_t + x_{it}'\beta_i - d_1) & \text{if } y = 0. \\ F(\alpha_i + \xi_t + x_{it}'\beta_i - d_1) - F(\alpha_i + \xi_t + x_{it}'\beta_i - d_2) & \text{if } y = 1. \\ F(\alpha_i + \xi_t + x_{it}'\beta_i - d_2) & \text{if } y = 2. \end{cases}$$

Example 3 (Count outcome) *When* $\mathcal{Y} = \{0, 1, 2, \ldots\}$*, model* (2.2.1) *becomes a count model with* $\sum_{y=0}^{\infty} g(y; v) = 1$ *for any* $v > 0$*. A leading example is Poisson count model:*

$$g(y; \alpha_i + \xi_t + x_{it}'\beta) = \frac{\exp(-\exp(\alpha_i + \xi_t + x_{it}'\beta))\exp(y(\alpha_i + \xi_t + x_{it}'\beta))}{y!}.$$

Another example of g *is negative binomial distribution.*

Example 4 (Multimodal outcome)

$$y_{it} = \underset{j=1,\ldots,J}{\arg\max} \left\{\alpha_{ij} + \xi_{tj} + x_{tj}'\beta_{ij} - u_{itj}\right\}, \tag{2.2.3}$$

where $(u_{it1}, \ldots, u_{itJ})$ *are independent of* $(\alpha_{ij}, \xi_{tj}, \beta_{ij}, x_{tj})_{j=1}^J$ *and distributed according to density* g^**. Define* $v_{itj} = \alpha_{ij} + \xi_{tj} + x_{tj}'\beta_{ij}$*. Then,*

$$g(y; v_{it1}, \ldots, v_{itJ}) = \sum_{j=1}^{J} \mathbf{1}\{y = j\}\mathbb{P}(u_{itj} - u_{itj'} \leq v_{itj} - v_{itj'}, \text{ for any } j' \neq j),$$

where the right-hand side is a function of g^* *and* J *indexes* $v_{it} = (v_{itj})_{j=1}^J$*. In this setting,* $\alpha_i = (\alpha_{ij})_{j=1}^J$*,* $\beta_i = (\beta_{ij})_{j=1}^J$*,* $\xi_t = (\xi_{tj})_{j=1}^J$*, and* J *is known.*

2.3 Identification and Estimation

Suppose that the econometrician observes (y_{it}, x_{it}) for $i \in \mathbf{N}$, and $t \in \mathbf{T}$ and aims to identify and estimate $(\alpha_i, \beta_i)_{i\in\mathbf{N}}$, $(\xi_t)_{i\in\mathbf{T}}$, and function g in model (2.2.1). To simplify the exposition, we present the arguments for $x_{it} = (x_{it}^{(1)}, x_{it}^{(2)}) \in \mathcal{X} \subset \mathbb{R}^2$

and the case of single index in the main text. The results for the case of multi-modal outcome (Example 4) and model (2.2.2) (heterogeneous slope parameters across time) are presented in Section 2.7.7 and 2.7.8, respectively. For any random variables U, Z, let $\text{Supp}(U|Z)$ denote the support of U conditional on Z. Without loss of generality, we normalize $\alpha_1 = 0$, $\xi_1 = 0$, and $\beta_1^{(1)} = 1$.[15]

2.3.1 Identification

In this section, we assume that both the number of individuals and that of time periods are infinity. To start with, define:

$$z_i(x^{(1)}; x^{(2)}) = \alpha_i + \beta_i^{(1)} x^{(1)} + x^{(2)}(\beta_i^{(2)} - \beta_1^{(2)}). \tag{2.3.1}$$

Intuitively, $z_i(x^{(1)}; x^{(2)})$ is interpreted as a *compensating variable*, i.e., the needed value of $x^{(1)}$ for individual 1 with $x^{(2)}$ to make her and i's indices equal: $\alpha_1 + \xi_t + \beta_1^{(1)} z_i(x^{(1)}; x^{(2)}) + \beta_1^{(2)} x^{(2)} = \alpha_i + \xi_t + \beta_i^{(1)} x^{(1)} + \beta_i^{(2)} x^{(2)}$. The following definition formalizes the idea of compensation.

Definition 2.3.1 (Compensable) *Individual i is compensable at $(x^{(1)}, x^{(2)}) \in \mathcal{X}_i = \text{Supp}(x_{it}|\alpha_i, \beta_i)$ by individual 1 if $(z_i(x^{(1)}; x^{(2)}) \in \mathcal{X}_1$.*[16]

Let $\mathcal{Z}_i = \left\{(z_i(x^{(1)}; x^{(2)}), x^{(2)}) : (x^{(1)}, x^{(2)}) \in \mathcal{X}_i\right\}$ and $\mathrm{P}_v(z, x^{(2)}) = v(z, x^{(2)})'$ be the operation of inner product. Denote by $\mathcal{Z}_{it}$ the support of $(z_i(x_{it}^{(1)}; x_{it}^{(2)}), x_{it}^{(2)})$ conditional on ξ_t. To obtain the main identification result, we propose the following assumption.

Assumption 2.3.1

(i). There exists $y \in \mathcal{Y}$ such that the function $g(y; v)$ is strictly monotonic in v.

(ii). (a) For all $i \in \mathbf{N}$, conditional on (α_i, β_i), $\{(y_{it}, x_{it})\}_{t \geq 2}$ is a strictly stationary and strong mixing process with mixing coefficients τ_t that satisfy $\tau_t \leq C\rho^t$. For all $t \in \mathbf{T}$, conditional on ξ_t, $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i \in \mathbf{N}}$ are independent.

(iii). For all $(i, i', t) \in \mathbf{N}^2 \times \mathbf{T}$, ξ_{it} and $(\alpha_i, \beta_i, x_{it}^{(1)})$ are independent conditional on $x_{it}^{(2)}$. Moreover, $\xi_t | \left\{x_{it}^{(2)} = x^{(2)}\right\} \overset{d}{=} \xi_t | \left\{x_{i't}^{(2)} = x^{(2)}\right\} \sim F_\xi(\xi; x^{(2)})$.

(iv). Any individual $i \in \mathbf{N}$ is compensable by individual 1 at least at $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$ for $k = 1, 2, 3$ with

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix}$$

being nonsingular.

[15] Suppose $\beta_1^{(1)} \neq 0$. Note that $g(y; \alpha_i + \xi_t + x_{it}'\beta_i) = \tilde{g}(y; \tilde{\alpha}_i + \tilde{\xi}_t + x_{it}'\tilde{\beta}_i)$, where $\tilde{\alpha}_i = (\alpha_i - \alpha_1)/\beta_1^{(1)}$, $\tilde{\xi}_t = (\xi_t - \xi_1)/\beta_1^{(1)}$, $\tilde{\beta}_i = \beta_i/\beta_1^{(1)}$, and $\tilde{g}(y; v) = g(y; \beta_1^{(1)} v + \alpha_1 + \xi_1)$. It is then necessary to normalize $\alpha_1 = 0$, $\xi_1 = 0$, and $\beta_1^{(1)} = 1$.

[16] We do not index $\mathcal{X}_i$ by t because $\{x_{it}\}_{t \in \mathbf{T}}$ will be assumed strictly stationary conditional on (α_i, β_i) for all $i \in \mathbf{N}$.

(v). Denote $\mathcal{Z}_t = \cap_{i\in\mathbf{N}}\mathcal{Z}_{it}$.

(a) For some $(t,r) \in \mathbf{T}\times\mathbb{R}$, $\left\{(z,x^{(2)}) \in \mathcal{Z}_t : \mathrm{P}_{(1,\beta_1^{(2)})}(z,x^{(2)}) = r\right\}$ *is not a singleton.*

(b) For all $t \in \mathbf{T}$, $\left(\cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it}) + \xi_t\right) \cap \left(\cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{i1})\right) \neq \emptyset$.

Remark 2.3.1 *When some covariates do not change across individuals, i.e.,* $x_{it} = x_t$*, we can condition on such covariates and* ξ_t *in Assumption 2.3.1(ii)b. Then, the independence among* $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i\in\mathbf{N}}$ *(with such covariates being excluded from* x_{it}*), as well as our main results below, still holds. See footnote 38 for more details.*

Assumption 2.3.1(i) is standard in index function models and satisfied in the examples we provided.[17] Assumption 2.3.1(ii) imposes dependence restrictions across individual and time dimensions of the panel. Assumption 2.3.1(ii)a requires stationarity and strong mixing properties across time. It allows ξ_t and $\xi_{t'}$ to be correlated, as long as the correlation vanishes as the time periods are distant enough. Assumption 2.3.1(ii)b requires the conditional cross-sectional independence across individuals. Both requirements are standard in the panel data literature.[18] Assumption 2.3.1(iii) requires the exogeneity of individual-fixed effect (α_i, β_i) and variable(s) $x_{it}^{(1)}$ with respect to time-fixed effect ξ_t, but allows $x_{it}^{(2)}$ to be endogenous, e.g., prices that are correlated with time-specific demand shocks.

Assumptions 2.3.1(iv)-(v) characterize the properties of the compensating variable $z_i(x^{(1)}; x^{(2)})$ that achieve the identification. Assumption 2.3.1(iv) specifies the condition under which $z_i(x^{(1)}; x^{(2)})$ compensates between individuals to identify individual-i specific parameters α_i, $\beta_i^{(1)}$, and $\beta_i^{(2)} - \beta_1^{(2)}$. Intuitively, if individual i can be compensated at three points in $\mathcal{X}_i$, one can then identify the three corresponding values of $z_i(x^{(1)}; x^{(2)})$ by comparing i's and 1's choices, y_{it} and y_{1t}, over time. The rank condition in Assumption 2.3.1(iv) ensures the unique recovery of α_i, $\beta_i^{(1)}$, and $\beta_i^{(2)} - \beta_1^{(2)}$ from the three identified values of $z_i(x^{(1)}; x^{(2)})$, achieving the identification. In essence, the rank requirement rules out the situation in which one point, say $(x^{(1)3}, x^{(2)3})$, lies on the line defined by $(x^{(1)1}, x^{(2)1})$ and $(x^{(1)2}, x^{(2)2})$. When $(x^{(1)}, x^{(2)})$ are continuous and the set of points at which i is compensable has positive Lebesgue measure, the rank condition automatically holds. Note that the compensation in Definition 2.3.1 can be generalized to any pair of two individuals i and i'. Instead of requiring individual 1 to compensate all the other individuals, one can also achieve the identification by relying on such pairwise compensation. In Appendix 2.7.1, we illustrate such identification argument and show that it may need weaker support conditions on $x_i^{(1)}$ than Assumption 2.3.1(iv) in some situations. Assumption 2.3.1(v)a gives the condition under which $z_i(x^{(1)}; x^{(2)})$ compensates between $x^{(1)}$ and $x^{(2)}$ for individual 1. It will be used to identify $\beta_1^{(2)}$ (and therefore $\beta_i^{(2)}$). Assumption

[17] For the case of multimodal outcome, we will replace the monotonicity by the conditions that imply the invertibility of g with respect to the vector of indices. See Assumption 2.7.2(i) for details.

[18] See Assumption 4.1 in Fernández-Val and Weidner (2016) for example.

2.3.1(v)b describes the condition under which $z_i(x^{(1)}; x^{(2)})$ compensates between time periods, i.e., the sets of indices in time periods t and 1 overlap. It will be used to identify (relative) time-specific fixed effect ξ_t. It is worth noting that when one is primarily interested in identifying the marginal effects of x_{it} on $\mathbb{P}(y_{it} = y|x_{i1}, \ldots, x_{it}, \alpha_i, \beta_i, \xi_t)$ in (2.2.1) rather than the values of $\beta_1^{(2)}$ and ξ_t, Assumption 2.3.1(v) can be redundant. In fact, if the unknown $g(y; v)$ is real analytic with respect to v for any $y \in \mathcal{Y}$ and $\mathcal{Z}_t$ contains an open subset, Assumption 2.3.1(iv) is already sufficient for this purpose.[19] We provide more details in Remark 2.7.1.

The support of $x_1^{(1)}$ (denoted by $\mathcal{X}_1^1$) plays an important role in the arguments of compensating variable and Assumptions 2.3.1(iv)-(v). When $x_1^{(1)}$ has a large support, e.g., $\mathcal{X}_1^1 = \mathbb{R}$, the two assumptions hold trivially. When $\mathcal{X}_1^1$ is not the entire real line (e.g., a box), both assumptions can still hold and the required support condition is determined by the ranges of (β_i, α_i), and ξ_t. We elaborate these points in the next example.

Example 5 (Support of $x_1^{(1)}$ and Identification) *Suppose that $\mathcal{X}_i = \mathcal{X} = [a, A] \times [b, B]$ where $a < A < 0$ and $0 < b < B$. Moreover, $\mathcal{Z}_{it} = \mathcal{Z}_i$ for any $(i, t) \in \mathbf{N} \times \mathbf{T}$. This setting can be considered as a demand model with $x^{(1)}$ being minus price of the goods and $x^{(2)}$ being its quality. Correspondingly, coefficient $\beta_i^{(1)} > 0$ (downward-sloping demand) is interpreted as the extent of the disutility of price and $\beta_i^{(2)}$ the preference for quality.*

In addition, suppose that $\max\{\beta_i^{(1)}\} > 1 > \min\{\beta_i^{(1)}\} > 1/\max\{\beta_i^{(1)}\} > 0$, $\frac{1}{2}\Delta_\beta^{(2)} := \max\{\beta_i^{(2)}\} - \beta_1^{(2)} = \beta_1^{(2)} - \min\{\beta_i^{(2)}\}$, and $\frac{1}{2}\Delta_\alpha := \max\{\alpha_i\} - \alpha_1 = \alpha_1 - \min\{\alpha_i\}$, i.e., individual 1's (α_1, β_1) is at the center of the range of $(\alpha_i, \beta_i) \in [\min\{\alpha_i\}, \max\{\alpha_i\}] \times [\min\{\beta_i^{(1)}\}, \max\{\beta_i^{(1)}\}] \times [\min\{\beta_i^{(2)}\}, \max\{\beta_i^{(2)}\}]$, where quantities defined by an application of the max *and* min *operators are well-defined. Recall the normalizations $\alpha_1 = 0$ and $\beta_1^{(1)} = 1$.*

First, Assumption 2.3.1(iv) holds when for any (α_i, β_i), there exists $x_i \in \mathcal{X}$ such that $\alpha_i + \beta_i^{(1)} x_i^{(1)} + x_i^{(2)}(\beta_i^{(2)} - \beta_1^{(2)}) \in (a, A)$. Because of the connectedness of $\mathcal{X}$, continuity of the linear mapping $x \to z_i(x^{(1)}; x^{(2)})$, and the intermediate value theorem, this is equivalent to

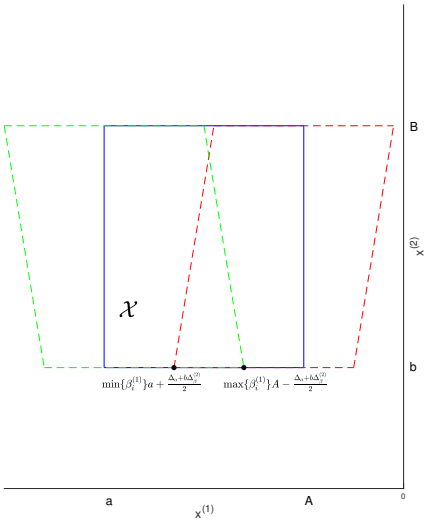
$$\begin{aligned}
&\sup_{(\alpha_i,\beta_i)} \inf_{x\in\mathcal{X}} \{\alpha_i + \beta_i^{(1)} x^{(1)} + x^{(2)}(\beta_i^{(2)} - \beta_1^{(2)})\} = \max\{\alpha_i\} + \min\{\beta_i^{(1)}\} a + b(\max\{\beta_i^{(2)}\} - \beta_1^{(2)}) < A, \\
&\inf_{(\alpha_i,\beta_i)} \sup_{x\in\mathcal{X}} \{\alpha_i + \beta_i^{(1)} x^{(1)} + x^{(2)}(\beta_i^{(2)} - \beta_1^{(2)})\} = \min\{\alpha_i\} + \max\{\beta_i^{(1)}\} A + b(\min\{\beta_i^{(2)}\} - \beta_1^{(2)}) > a.
\end{aligned} \tag{2.3.2}$$

The geometric interpretation of (2.3.2) *is illustrated in Figure 2.1(a). The linear mapping $x \to (z_i(x^{(1)}; x^{(2)}), x^{(2)})$ maps the box $\mathcal{X}$ to a parallelogram that overlaps with int($\mathcal{X}$), the interior of $\mathcal{X}$ (e.g., the red and green ones in Figure 2.1(a)). The first inequality in* (2.3.2) *requires the red parallelogram corresponding to the mapping defined by $(\max\{\alpha_i\}, \min\{\beta_i^{(1)}\}, \max\{\beta_i^{(2)}\})$, which is stretched to the right, to overlap with*

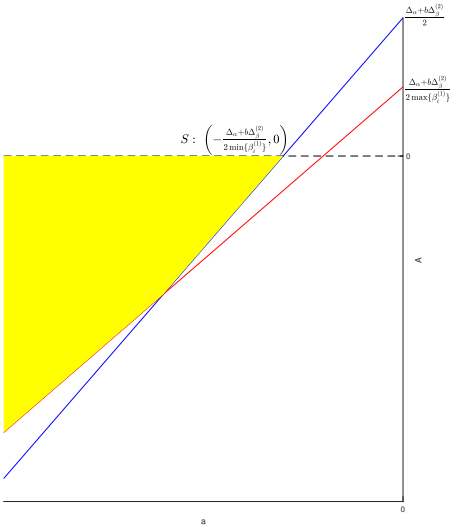
[19] A function $g(v)$ is real analytic at $v = v_0$ if $g(v)$ is C^∞ and coincides with its Taylor series (defined around v_0) in a neighborhood of v_0. Most link functions, e.g., logit, probit, multinomial logit, are real analytic. Iaria and Wang (2022) also show that mixed-logit/probit models with an index structure as in (2.2.1) are also real analytic.

FIGURE 2.1: Support Condition on $x^{(1)}$ and Identification

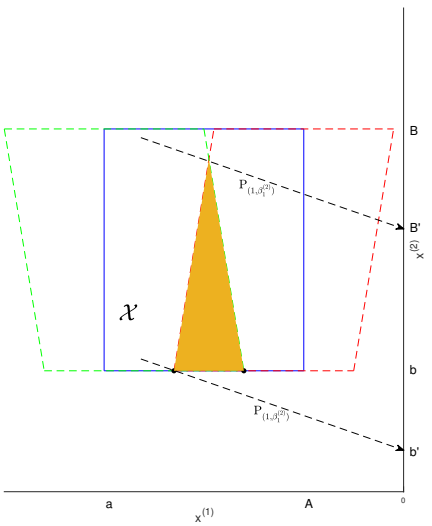
(A) Graphic illustration of Assumption 2.3.1(iv)



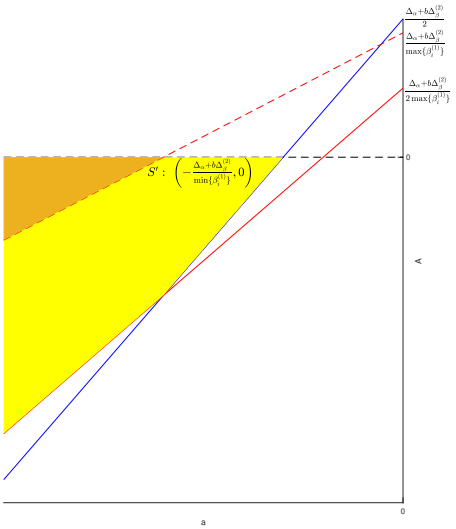
(B) (a, A) with which Assumption 2.3.1(iv) holds



(C) Graphic illustration of Assumption 2.3.1(v)



(D) (a, A) with which Assumption 2.3.1(v)a holds



int($\mathcal{X}$). *Similarly, the second inequality requires the green parallelogram corresponding to* $(\min\{\alpha_i\}, \max\{\beta_i^{(1)}\}, \min\{\beta_i^{(2)}\})$, *which is stretched to the left, to overlap with int*($\mathcal{X}$). *These two parallelograms are the "most distant" from* $\mathcal{X}$ *in either direction. Inequalities* (2.3.2) *are further equivalent to:*

$$\begin{aligned} A &> \min\{\beta_i^{(1)}\}a + \frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2}, \\ A &> \frac{1}{\max\{\beta_i^{(1)}\}}a + \frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2\max\{\beta_i^{(1)}\}}. \end{aligned} \tag{2.3.3}$$

The yellow region defined by the blue line (the first inequality in (2.3.3)*) and red line (the second in* (2.3.3)*) in Figure 2.1(b) shows the values of* (a, A) *that satisfy* (2.3.3). *In particular,* A *can be close to zero (the lower bound of the observed price of the goods is small) and* a *close to* $-\frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2\min\{\beta_i^{(1)}\}}$, *as illustrated by point* S. *The size of the corresponding support for* $x^{(1)}$, $A - a$, *is then close to* $\frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2\min\{\beta_i^{(1)}\}}$. *For any size greater than this length, Assumption 2.3.1(iv) can always hold with some* (a, A). *This minimal support requirement becomes more stringent when the ranges of* α_i (Δ_α) *and* $\beta_i^{(2)}$ $(\Delta_\beta^{(2)})$ *increase.*

Second, Assumption 2.3.1(v)a holds when one further requires that the parallelograms the most distant from $\mathcal{X}$ *overlap, as illustrated by the orange region in Figure 2.1(c). This is because the preimage of* $\mathrm{P}_{(1,\beta_1^{(2)})}(x) = r$ *for any* $r \in (b', B')$ *is a line segment in this region and therefore not a singleton. In particular, this implies*

$$\begin{aligned} &\max\{\alpha_i\} + \min\{\beta_i^{(1)}\}a + b(\max\{\beta_i^{(2)}\} - \beta_1^{(2)}) < \min\{\alpha_i\} + \max\{\beta_i^{(1)}\}A + b(\min\{\beta_i^{(2)}\} - \beta_1^{(2)}) \\ \implies &A > \frac{\min\{\beta_i^{(1)}\}}{\max\{\beta_i^{(1)}\}}a + \frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{\max\{\beta_i^{(1)}\}}. \end{aligned} \tag{2.3.4}$$

Inequality (2.3.4) *is stronger than the second one in* (2.3.3) *and is represented by the dashed red line in Figure 2.1(d). The values of* (a, A) *with which Assumption 2.3.1(v)a holds, the orange region in Figure 2.1(d), are then more limited than the yellow one (corresponding to Assumption 2.3.1(iv)) and the strict lower bound of* $A - a$, $\frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{\min\{\beta_i^{(1)}\}}$ *(achieved at* S'*), is greater than* $\frac{\Delta_\alpha + b\Delta_\beta^{(2)}}{2\min\{\beta_i^{(1)}\}}$. *In other words, identifying further* $\beta_1^{(2)}$ *requires a larger support of* $x^{(1)}$ *than the one needed for the identification of* α_i, $\beta_i^{(1)}$, *and* $\beta_i^{(2)} - \beta_1^{(2)}$.

Finally, Assumption 2.3.1(v)b holds when the projection of the orange region by $\mathrm{P}_{(1,\beta_1^{(2)})}$ *(the segment between* b' *and* B' *in Figure 2.1(c)) intersects with itself when translated by* ξ_t *for any* $t \in \mathbf{T}$. *Without enlarging the support of* $x^{(1)}$ *required by Assumption 2.3.1(v)(a), one can only identify* ξ_t *with* $|\xi_t| < B' - b'$. *As a result, to point identify* ξ_t *with* $|\xi_t| \geq B' - b'$, *one may need a larger support of* $x^{(1)}$ *than that required by Assumption 2.3.1(v)a.*

The next theorem summarizes our main identification result. See Appendix 2.7.1 for the proof.

Theorem 2.3.1 *Suppose that Assumptions 2.3.1(i)-(iv) hold.*

- $\beta_i^{(1)}$, α_i, *and* $\beta_i^{(2)} - \beta_1^{(2)}$ *are identified for* $i \in \mathbf{N}$.
- *If Assumptions 2.3.1(v) further holds, then*
  - ξ_t *and* $\beta_i^{(2)}$ *are identified for* $i \in \mathbf{N}$ *and* $t \in \mathbf{T}$.
  - $g(y; v)$ *is identified for* $(y, v) \in \mathcal{Y} \times \cup_{t \in \mathbf{T}} \left(\cap_{i \in \mathbf{N}} \mathrm{P}_{(1, \beta_1^{(2)})}(\mathcal{Z}_{it}) + \xi_t \right)$.

According to Theorem 2.3.1, fixed effects parameters α_i, β_i, and ξ_t are point identified when N and T are large. In particular, the identification of β_i enables applied researchers to specify and estimate unobserved heterogeneity in the causal effects of covariates among individuals. This is important when the researcher is interested in the distributional effects of such covariates and their policy implications. In Section 2.5, we illustrate this point by revisiting two classic empirical studies. Moreover, Theorem 2.3.1 provides a theoretical foundation for nonparametrically estimating the link function g. One such procedure is sieve MLE (see Chen et al. (2006); Gallant and Nychka (1987); Shen and Wong (1994) for examples), which can be applied in practice to check whether empirical findings are driven by parametric assumptions such as logit and probit often motivated by computational reasons. Finally, Theorem 2.3.1 can be extended to a model (2.2.1) with multimodal outcomes and model (2.2.2) with heterogeneous slopes β_t. See Sections 2.7.7 and 2.7.8 for details, respectively.

2.3.2 Estimation

In this section, we propose a convenient iterative estimation procedure of model (2.2.1). It has three appealing features. First, it significantly improves the numerical efficiency upon the routine implementation of the MLE, particularly when one (or both) dimension in the panel is large. Second, we show that it is numerically equivalent to the MLE under standard conditions: the resulting estimators converge to the MLE estimator as long as the number of iterations is large enough. Finally, the proposed estimation procedure is of general interest. It applies to both finite-T and large-T settings with a post-estimation bias reduction and correction, respectively. We present a semi-parametric estimation of model (2.2.1) with a known g and $\beta_i = \beta$ in the main text. We discuss the extensions to the settings in which one estimates g and/or heterogeneous slope parameters in Remarks 2.3.2 and 2.3.3.

Oftentimes, researchers estimate model (2.2.1) by treating the unobserved individual and time effects as parameters to be estimated and using a concentrated MLE. Denote the log-likelihood function by

$$\mathscr{L}_{NT}(\theta) := \sum_{i=1}^{N} \sum_{t=1}^{T} \log g(y_{it}; \alpha_i + \xi_t + x_{it}'\beta), \tag{2.3.5}$$

where $\theta = (\alpha_2, \ldots, \alpha_N, \xi_1, \ldots, \xi_T, \beta)$ with α_1 being normalized to zero. The standard implementation consists of two steps. In the first step (inner loop), given β, one maximizes $\mathscr{L}_{NT}(\theta)$ with respect to fixed effect parameters (α_i, ξ_t) for $i = 2, \ldots, N$ and $t = 1, \ldots, T$:

$$(\widehat{\alpha}_2, \ldots, \widehat{\alpha}_N, \widehat{\xi}_1, \ldots, \widehat{\xi}_T) \in \underset{(\alpha_2,\ldots,\alpha_N)\in\mathcal{A},(\xi_1,\ldots,\xi_T)\in\Xi}{\arg\max} \mathscr{L}_{NT}(\alpha_2, \ldots, \alpha_N, \xi_1, \ldots, \xi_T, \beta), \tag{2.3.6}$$

where $\mathcal{A} := \mathcal{A}_2 \times \cdots \times \mathcal{A}_N \subset \mathbb{R}^{N-1}$ and $\Xi := \Xi_1 \times \cdots \times \Xi_T \subseteq \mathbb{R}^T$, with $\mathcal{A}_i$ and Ξ_t containing the support of α_i and ξ_t, respectively. In the second step (outer loop), plugging in the estimates of the fixed effects in (2.3.5), one maximizes $\mathscr{L}_{NT}(\theta)$ with respect to β:

$$\widehat{\beta} \in \underset{\beta\in\mathcal{B}}{\arg\max}\ \mathscr{L}_{NT}(\widehat{\alpha}_2, \ldots, \widehat{\alpha}_N, \widehat{\xi}_1, \ldots, \widehat{\xi}_T, \beta), \tag{2.3.7}$$

where $\mathcal{B} \subset \mathbb{R}^K, K \geq 1$.

This standard implementation can be computationally intensive due to two reasons. First, concentration step (2.3.6) involves numerical optimization with a large number of parameters (i.e., fixed effects whose number is at least of order $T + N$). Simultaneous numeric searches with respect to these parameters are both time and space consuming. Second, and more severely, the maximization in outer loop (2.3.7) treats $\widehat{\alpha}_i$ and $\widehat{\xi}_t$ as functions of β (as a result of the inner loop). Each numeric search in this step will then inevitably execute (2.3.6) multiple times, substantially increasing computation time.

Our proposed iterative procedure resembles the block-nonlinear Gauss-Seidel method (or the bloc/cyclic coordinate descent method) in the optimization literature (see, e.g., Bertsekas, 2016) and circumvents these two numerical challenges in the implementation of the likelihood estimators. In each iteration, we update sequentially estimated individual fixed effects $\{\widehat{\alpha}_i\}_{i=2}^N$, time fixed effects $\left\{\widehat{\xi}_t\right\}_{t=1}^T$ and common slopes $\widehat{\beta}$. In particular, the updates of $\{\widehat{\alpha}_i\}_{i=2}^N$ and $\left\{\widehat{\xi}_t\right\}_{t=1}^T$ are fully parallelized, greatly reducing computational time and solving the numerical challenge in concentration step (2.3.6). This is doable due to the two-way fixed effects structure in (2.3.6): given $\{\xi_t\}_{t=1}^T$ and β, when maximizing the entire likelihood with respect to α_i, only the likelihood corresponding to individual i is relevant. Then, to update $\{\widehat{\alpha}_i\}_{i=2}^N$, one only needs to solve $N-1$ one-dimensional maximization problems in parallel. Analogously, given $\{\alpha_i\}_{i=2}^N$ and β, when maximizing the entire likelihood with respect to ξ_t, only the likelihood corresponding to time t is relevant. Then, to update $\left\{\widehat{\xi}_t\right\}_{t=1}^T$, one only needs to solve T one-dimensional maximization problems in parallel given the updated $\{\widehat{\alpha}_i\}_{i=2}^N$ and $\widehat{\beta}$. Finally, given the updated $\{\widehat{\alpha}_i\}_{i=2}^N$ and $\left\{\widehat{\xi}_t\right\}_{t=1}^T$, we update $\widehat{\beta}$. This update avoids re-evaluating $\{\widehat{\alpha}_i\}_{i=2}^N$ and $\left\{\widehat{\xi}_t\right\}_{t=1}^T$, solving the numerical challenge in (2.3.7).

We provide two algorithms that practitioners can use depending on the dimensionality of the optimization and availability of computational resources. The first one, "fixed-point MLE" (FPMLE), updates $\{\widehat{\alpha}_i\}_{i=2}^N$, $\left\{\widehat{\xi}_t\right\}_{t=1}^T$ and $\widehat{\beta}$ by solving the corresponding optimization problems in each iteration.

Algorithm FPMLE:

1. Let $(\xi_1^{(0)}, \ldots, \xi_T^{(0)}, (\beta^{(0)})')' \in \Xi \times \mathcal{B}$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, \ldots\}$. Set $s = 0$.

2 Compute (in parallel) for all $i \in \{2, \ldots, N\}$:

$$\alpha_i^{(s+1)} \in \arg\max_{\alpha \in \mathcal{A}_i} \sum_{t=1}^T \log g(y_{it}; \alpha + x_{it}'\beta^{(s)} + \xi_t^{(s)}).$$

3. Compute (in parallel) for all $t \in \{1, \ldots, T\}$:

$$\xi_t^{(s+1)} \in \arg\max_{\xi \in \Xi_t} \sum_{i=1}^N \log g(y_{it}; \alpha_i^{(s+1)} + x_{it}'\beta^{(s)} + \xi).$$

4. Compute:

$$\beta^{(s+1)} \in \arg\max_{\beta \in \mathcal{B}} \sum_{i=1}^N \sum_{t=1}^T \log g(y_{it}; \alpha_i^{(s+1)} + x_{it}'\beta + \xi_t^{(s+1)}).$$

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

When N (or T) is large, or computational resource is limited (e.g., the number of CPUs available to parallel computation), Steps 2 and 3 in FPMLE could still be time-consuming. This motivates our second algorithm, an accelerated version of FPMLE, labelled as FPMLE^{++}, that updates $\{\widehat{\alpha}_i\}_{i=2}^N$, $\left\{\widehat{\xi}_t\right\}_{t=1}^T$, and $\widehat{\beta}$ using one-step Newton-Raphson method, rather than solving the optimization problems. Let $g'(y; v)$ denotes the first derivative of g with respect to its second argument.

Algorithm FPMLE^{++}:

1. Let $(\alpha_2^{(0)}, \ldots, \alpha_N^{(0)}, \xi_1^{(0)}, \ldots, \xi_T^{(0)}, (\beta^{(0)})')' \in \mathcal{A} \times \Xi \times \mathcal{B}$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, \ldots\}$. Let $\left\{\nu^{(s)}\right\}_{s \geq 0}$ be some bounded sequence of positive scalars such that $\liminf_s \nu^{(s)} > 0$. Set $s = 0$.

2 Compute:

$$\begin{pmatrix} \alpha_2^{(s+1)} \\ \vdots \\ \alpha_N^{(s+1)} \end{pmatrix} = \left[\begin{pmatrix} \alpha_2^{(s)} \\ \vdots \\ \alpha_N^{(s)} \end{pmatrix} - \nu^{(s)} \begin{pmatrix} \sum_{t=1}^T \frac{g'}{g}(y_{2t}; \alpha_2^{(s)} + x_{2t}'\beta^{(s)} + \xi_t^{(s)}) \\ \vdots \\ \sum_{t=1}^T \frac{g'}{g}(y_{Nt}; \alpha_N^{(s)} + x_{Nt}'\beta^{(s)} + \xi_t^{(s)}) \end{pmatrix} \right]_{\mathcal{A}}^+,$$

where $[v]_{\mathcal{A}}^+$ denotes the vector whose i-th coordinate is the orthogonal projection of v_i on $\mathcal{A}_i$.

3. Compute:

$$\begin{pmatrix} \xi_1^{(s+1)} \\ \vdots \\ \xi_T^{(s+1)} \end{pmatrix} = \left[\begin{pmatrix} \xi_1^{(s)} \\ \vdots \\ \xi_T^{(s)} \end{pmatrix} - \nu^{(s)} \begin{pmatrix} \sum_{i=1}^N \frac{g'}{g}(y_{i1}; \alpha_i^{(s+1)} + x_{i1}'\beta^{(s)} + \xi_1^{(s)}) \\ \vdots \\ \sum_{i=1}^N \frac{g'}{g}(y_{iT}; \alpha_i^{(s+1)} + x_{iT}'\beta^{(s)} + \xi_T^{(s)}) \end{pmatrix} \right]_\Xi^+,$$

where $[v]_\Xi^+$ denotes the vector whose t-th coordinate is the orthogonal projection of v_t on Ξ_t.

4. Compute:

$$\beta^{(s+1)} = \left[\beta^{(s)} - \nu^{(s)} \sum_{i=1}^N \sum_{t=1}^T x_{it} \frac{g'}{g}(y_{it}; \alpha_i^{(s+1)} + x_{it}'\beta^{(s)} + \xi_t^{(s+1)}) \right]_\mathcal{B}^+,$$

where $[v]_\mathcal{B}^+$ denotes the orthogonal projection of v on $\mathcal{B}$.

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

Note that in Step 2 (and 3), the update of $\alpha_i^{(s)}$ ($\xi_t^{(s)}$) is purely arithmetic. In particular, when $\mathcal{A}_i$ (and Ξ_t) are "nice" convex sets, e.g., boxes, the update does not involve any $\alpha_r^{(s)}$ for $r \neq i$ ($\xi_r^{(s)}$ for $r \neq t$). As a result, these updates can be entirely vectorized within each step. While the parallelization in FPMLE is usually constrained by the number of CPUs, the vectorization in FPMLE^{++} is not and can be implemented on the GPUs, further accelerating the implementation.

Remark 2.3.2 (Estimating heterogeneous slope parameters) *The extension of FPMLE/FPMLE^{++} to the case of heterogeneous slopes β_i (β_t) is straightforward. It suffices to additionally update $\beta_i^{(s)}$ in Step 2 ($\beta_t^{(s)}$ in Step 3) using the same rule in either algorithm. We provide more details in Sections 2.7.9.*

Remark 2.3.3 (Estimating link function g) *Denote by $\mathcal{G}$ the space of a finite-dimensional parameters, θ^g, that determine the link function g. We use $g(\cdot, \theta^g)$ to refer to the parametrizaton by $\theta^g \in \mathcal{G}$. For instance, g can be a link function corresponding to t-distribution with θ^g being the degree of freedom of the t-distribution. Another example is sieve estimation of g. In the setting of binary outcome (Example 1), for given N and T, $\mathcal{G}$ can be a finite-dimensional space with the density of g, $den(g)$, satisfies:*

$$den(g)(\delta; \theta^g) = \left[\sum_{k=1}^K \theta_k^g H_k(x) \exp\left\{ -\frac{x^2}{2} \right\} \right]^2, \tag{2.3.8}$$

where H_k is the Hermite polynomial of degree k. $\mathcal{G}$ can also be defined as a mixture sieve space with

$$den(g)(\delta; \theta^g) = \sum_{k=1}^K \theta_k^g \frac{1}{\sigma_k} \phi\left(\frac{x - \mu_k}{\sigma_k} \right), \tag{2.3.9}$$

where ϕ is the standard normal density.

To customize FPMLE and FPMLE^{++} to estimating θ^g, it suffices to add an additional step between Steps 4 and 5 in each iteration of FPMLE and FPMLE^{++} that updates the estimated parameters $\theta^g \in \mathcal{G}$. In FPMLE, this step is

- *Compute:*

$$\theta^{g(s+1)} \in \arg\max_{\theta \in \mathcal{G}} \sum_{i=1}^{N}\sum_{t=1}^{T} \log g(y_{it}; \alpha_i^{(s+1)} + x_{it}'\widehat{\beta}^{(s+1)} + \xi_t^{(s+1)}, \theta).$$

In FPMLE^{++}, this step is

- *Compute:*

$$\theta^{g(s+1)} = \left[\theta^{g(s)} - \nu^{(s)} \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it} \frac{\partial_\theta g}{g}(y_{it}; \alpha_i^{(s+1)} + x_{it}'\beta^{(s+1)} + \xi_t^{(s+1)}, \theta^{g(s)})\right]_{\mathcal{G}}^{+},$$

  where $[v]_{\mathcal{G}}^{+}$ denotes the orthogonal projection of v on $\mathcal{G}$.

Remark 2.3.4 *In theory, the projection operation in each step of FPMLE^{++} ensures that the updated estimate is always within its support, facilitating the numerical convergence by restricting the estimates in the first iterations to be not too distant from the solutions. In practice, one can update the estimates without projecting and still achieve the numerical convergence, which we observe in the Monte Carlo simulations.*

2.3.3 Numerical Equivalence to the MLE

In this section, we prove that both FPMLE and FPMLE^{++} are numerically equivalent to the MLE: given T and N, both estimators converge to the MLE estimators as the number of iterations increases to infinity. As a result, FPMLE and FPMLE^{++} provide reliable approximations to the MLE in the finite sample.

Define the MLE, $\widehat{\theta}_{NT}^{\text{MLE}}$, that maximizes the log-likelihood function $\mathscr{L}_{NT}(\cdot)$ over $\mathbb{R}^{K+N+T-1}$:

$$\widehat{\theta}_{NT}^{\text{MLE}} = \arg\max_{\theta \in \mathbb{R}^{K+N+T-1}} \mathscr{L}_{NT}(\theta). \tag{2.3.10}$$

To establish the equivalence results, we will need the following assumptions on $\mathscr{L}_{NT}(\cdot)$ and $\Theta_{NT} := \mathcal{A} \times \Xi \times \mathcal{B}$.

Assumption 2.3.2

(i). $\mathcal{A}_i, \Xi_t$, and $\mathcal{B}$ are convex, closed sets with nonempty interior and Θ_{NT} contains $\widehat{\theta}_{NT}^{\text{MLE}}$.

(ii). $\mathscr{L}_{NT}$ is strictly concave and continuously differentiable over $\mathbb{R}^{K+N+T-1}$. Moreover, $\lim_{\|\theta\|\to\infty} \mathscr{L}_{NT}(\theta) = -\infty$.

(iii). $\mathcal{A}_i, \Xi_t$, and $\mathcal{B}$ are bounded boxes containing 0 in their interiors and $\mathscr{L}_{NT}$ is twice continuously differentiable over over $\mathbb{R}^{K+N+T-1}$.

Assumption 2.3.2(i) is a standard condition for deriving properties of M-estimators. Assumption 2.3.2(ii) features smoothness, concavity and coercivity properties of $\mathscr{L}_{NT}$ that together ensure that problem (2.3.10) admits a unique solution characterized by the first-order conditions. The commonly used nonlinear models in applied economics such as logit, probit, ordered probit, Poisson, and tobit models satisfy this assumption, provided that all the elements of x_{it} have sufficient cross sectional and time series variation.

Assumptions 2.3.2(i)-(ii) are sufficient for the numerical equivalence of FPMLE. We further need Assumption 2.3.2(iii) to show the numerical equivalence of FPMLE^{++}. This assumption strengthens the smoothness of the log-likelihood function and holds generically for commonly used distributions. Together with Assumptions 2.3.2(i)-(ii), it allows to locally bound the Hessian matrix of $\mathscr{L}_{NT}(\cdot)$ from below and from above. This will deliver local strong concavity and Lipschitzity of the gradient, which is the key to the convergence of FPMLE^{++}.

We now state the numerical equivalence of FPMLE and FPMLE^{++} to the MLE. The proof can be found in Appendix 2.7.2.

Theorem 2.3.2

- *Suppose that Assumptions 2.3.2(i)-(ii) hold. Then, $\widehat{\theta}_{NT}^{\text{MLE}}$ exists and the sequence of iterates generated by FPMLE, $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,\ldots}$, converges to $\widehat{\theta}_{NT}^{\text{MLE}}$.*
- *If Assumption 2.3.2(iii) further holds and $\nu^{(s)} \equiv \nu$ is constant such that $0 < \nu < 1/\bar{L}$ for some absolute constant $\bar{L} > 0$,[20] then the sequence of iterates generated by FPMLE^{++}, $\left\{\widehat{\theta}_{NT}^{++(s)}\right\}_{s=1,2,\ldots}$, converges to $\widehat{\theta}_{NT}^{\text{MLE}}$.*

Remark 2.3.5 (Estimating heterogeneous slope parameters) *The numerical convergences of both algorithms still hold in the presence of heterogeneous slopes β_i. We refer to the proof in Appendix 2.7.2 for such extension.*

When the concavity requirement in Assumption 2.3.2(ii) does not hold, the likelihood function may have multiple local maxima and the numerical equivalence is not universally guaranteed in theory. The lack of concavity is more likely to occur when the link function g is left to be estimated.[21] In this case, one can still verify the numerical convergence of both algorithms by simply checking if $\left\|\widehat{\theta}_{NT}^{(s+1)} - \widehat{\theta}_{NT}^{(s)}\right\|$ and/or the corresponding difference in the likelihood function is small enough. In Section 2.7.9, we show that if FPMLE/FPMLE^{++} converges numerically, it then converges to a

[20] For the definition of $\nu^{(s)}$, see Algorithm FPMLE^{++}. The constant $\bar{L}$ is implicitely defined in the proof of Theorem 2.3.2, Eq. (2.7.10). In practice, choosing ν and deriving an upper bound on $\bar{L}$ is straightforward given knowledge of g, Θ_{NT}, and the data.

[21] For instance, in the case of ordinal outcome, the log-concavity of $\text{den}(g)$ is sufficient (and almost necessary) for the concavity of $\mathscr{L}_{NT}$ with respect to θ (Pratt, 1981). When $\text{den}(g)$ is left to be estimated, e.g., the sieve approach in (2.3.8) and (2.3.9), such shape restriction on $\text{den}(g)$ is, a priori, not imposed. Even with such restrictions, one still needs the concavity of $\mathscr{L}_{NT}$ with respect to the entire vector of parameters, i.e., θ and those in $\text{den}(g)$, to obtain the numerical equivalence in Theorem 2.3.2, which is not automatically satisfied.

stationary point of the likelihood function (2.3.5). As a result, applied researchers can use both algorithms to pin down the set of stationary points of the likelihood function using different starting points, and obtain the global maximum from the set. Theorem 2.3.2 implies that given (N,T), $\widehat{\theta}_{NT}^{(s)}$ and $\widehat{\theta}_{NT}^{++(s)}$ will be close enough to $\widehat{\theta}_{NT}^{\text{MLE}}$ when s is large. In our Monte Carlo simulations (Section 2.4) as well as empirical illustrations, both algorithms converge fast and already achieve good numerical approximation after a few number of iterations.

Inference. Given the numerical equivalence, the researcher can use $\widehat{\theta}_{NT}^{(s)}$ and $\widehat{\theta}_{NT}^{++(s)}$ as approximates of $\widehat{\theta}_{NT}^{\text{MLE}}$ and conduct inference.[22] In the classic setting with $\beta_i = \beta$, one can implement post-estimation bias correction by deriving a consistent estimate of the bias (Fernández-Val and Weidner, 2016) and bias reduction using re-sampling methods such as jackknife (Dhaene and Jochmans, 2015; Hahn and Newey, 2004). In the setting with heterogeneous β_i, Boneva and Linton (2017) propose a method of inference when T and N are both large with $T/N \to 0$ (see Assumption B2 on page 1230). Gao et al. (2020) focus on the inference in a binary panel model with heterogeneous slopes, interactive fixed effects, and a known link function. To be self-contained, we provide a consistency result for the MLE estimators of the slopes in Appendix 2.7.3, and show that $\max_{i=1}^{N} |\widehat{\beta}_i - \beta_i^0|$ is of order $N^{-3/8}$ as $N \to \infty$ and $N/T \to \kappa \in (0, +\infty)$ and therefore $(\widehat{\beta}_i)_{i=1}^{N}$ is consistent under the max norm. In particular, this result implies that plug-in estimators of the moments of β_i are consistent. Consequently, applied researchers can estimate the population average of the causal effect of a covariate (the mean of β_i) and assess the extent of its heterogeneity across individuals (the dispersion of β_i). We also provide Monte Carlo evidence for the consistency result in the setting of Poisson count model in Table 2.8 of Section 2.7.11. In the next section, we supplement the consistency result with a practical Bootstrap inference procedure and provide Monte Carlo evidence for its good finite-sample performance. [23]

2.4 Monte Carlo Experiments

In this section, we use Monte Carlo experiments to assess the numerical performance of FPMLE and FPMLE^{++}. We focus on three tasks. First, we investigate the number of iterations with which the objects of interests, e.g., slope parameters, average partial effects (APEs), computed by using $\widehat{\theta}_{NT}^{(s)}$ and $\widehat{\theta}_{NT}^{++(s)}$ approximate well those obtained

[22] We focus on the inference in the semi-parametric setting of model (2.2.1) with g being known (or up to a fixed number of parameters that does not increase with N and T). We leave the inference in the nonparametric setting, e.g., sieve estimation of g in (2.3.8) and (2.3.9), for future research.

[23] As detailed in Appendix 2.7.3, when β_i is bounded, we show that the plug-in estimators of the moments of β_i converge to the true values with a rate at least equal to $N^{3/8}$. Despite the good finite-sample performance of the proposed Bootstrap procedure suggested by the Monte Carlo simulations, it is yet to show the theoretical validity of the procedure for such plug-in estimators (e.g., they are $\sqrt{N}$-Gaussian, potentially subject to an asymptotic bias). We leave this for future research. An alternative inference method is subsampling (Politis et al., 1999) that applies under weaker conditions than Bootstrap (e.g., a convergence rate slower than $\sqrt{N}$).

by using $\widehat{\theta}^{\text{MLE}}_{NT}$. Second, we investigate the extent to which both algorithms reduce execution time relative to the routine implementation of the MLE, and in particular, the performance of FPMLE^{++} when the number of fixed effects is large. Third, in the presence of heterogeneous slopes, we assess the finite-sample performance of a practical Bootstrap inference procedure for the distributional features of the slopes and the APEs. Finally, based on our findings, we give practical guidance to applied researchers regarding the use of our algorithms.

Monte Carlo design. The designs build on those in Section 5.1 of Fernández-Val and Weidner (2016). We consider a static logit model with homogeneous slope coefficients (Example 1):

$$y_{it} = \mathbf{1}\{x_{it}\beta_0 + \alpha_i + \xi_t \geq u_{it}\}, \quad i = 1, \dots, N,\, t = 1, \dots, T,$$

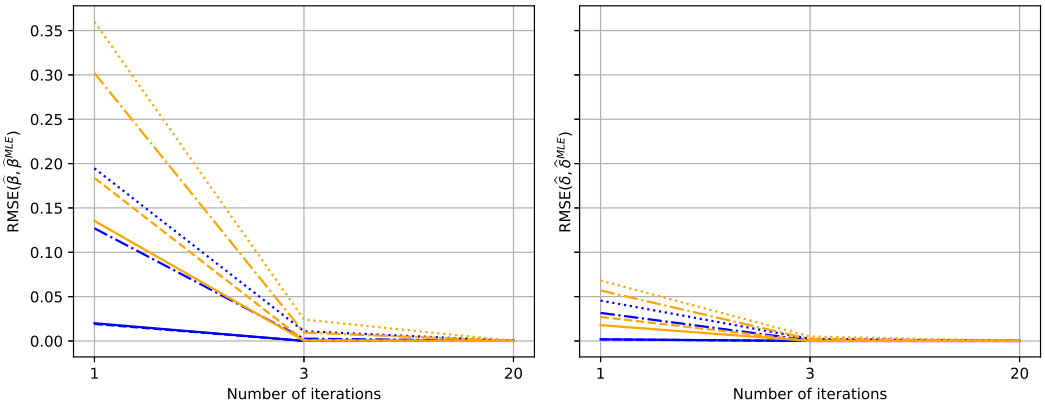
where $\alpha_1 = 0, \alpha_i \sim \mathcal{N}(0, 1/16)$ for $i \geq 2$, $\xi_t \sim \mathcal{N}(0, 1/16)$, $u_{it} \sim \Lambda$ with $\Lambda(u) = 1/(1 + \exp(-u))$, and $\beta_0 = 1$. In all designs x_{it} is strictly exogenous with respect to u_{it} conditional on the individual and time effects. The variables $\alpha_i, \xi_t, u_{it}, v_{it}$, and x_{i0} are independent and i.i.d. across individuals and time periods. We consider four data generating processes (DGPs) for x_{it}. In DGP (i), $x_{it} \sim \mathcal{N}(0, 1)$. In DGP (ii), $x_{it} \sim$Unif$[-\sqrt{3}, \sqrt{3}]$. Both DGPs satisfy Assumption 2.3.1. In DGP (iii), $x_{it} = x_{i,t-1}/2 + \alpha_i + \xi_t + v_{it}$, with $v_{it} \sim \mathcal{N}(0, 1/2)$, and $x_{i0} \sim \mathcal{N}(0, 1)$. In DGP (iv), $x_{it} = 2t/T + \alpha_i + \xi_t + v_{it}$, with $v_{it} \sim \mathcal{N}(0, 3/4)$. DGPs (iii) and (iv) violate the exogeneity condition (Assumption 2.3.1(iii)). Besides, DGP (iv) violates the stationary requirement in Assumption 2.3.1(ii)a.

Calibration of the number of iterations. Figure 2.2 summarizes the "Root Mean Squared Error" (RMSE) distance to the MLE estimator for $\widehat{\beta}_{NT}$ (blue, left panel), $\widehat{\beta}^{++}_{NT}$ (orange, left panel), and the corresponding estimated APEs $\widehat{\delta}_{NT}$ and $\widehat{\delta}^{++}_{NT}$ (right panel).[24] The statistics in both figures are computed using 50 Monte Carlo replications and with $N = T = 200$. The full results are in Table 2.2.

As the number of iterations increases, both FPMLE and FPMLE^{++} converge to the MLE as predicted by Theorem 2.3.2. FPMLE delivers good approximation to the MLE even when the number of iteration is small. When we run only 3 iterations, the RMSE distance for $\widehat{\beta}_{NT}$ is at most of order 10^{-3} for all the DGPs and the RMSE distance for $\widehat{\delta}_{NT}$ is even a magnitude smaller. For a given number of iterations, FPMLE^{++} approximates less well than FPMLE. However, the difference seems to diminish very quickly. For the DGPs we consider, FPMLE^{++} with 20 iterations already achieves comparable precision to FPMLE (see Table 2.2). We replicate the exercises for the setting of $N \gg T$ ($N = 5000$ and $T = 30$) and these findings remain valid. For details, see Table 2.7 in Section 2.7.11.

[24] For the definitions of the RMSE distance to the MLE estimator $\widehat{\theta}^{\text{MLE}}$ and $\widehat{\delta}_{NT}$, see Eq. (2.7.20) and (2.7.21), respectively.

FIGURE 2.2: Numerical Convergence of $\widehat{\beta}_{NT}$, $\widehat{\beta}^{++}_{NT}$, $\widehat{\delta}_{NT}$, and $\widehat{\delta}^{++}_{NT}$



Notes: $N = T = 200$. Left panel: $\widehat{\beta}_{NT}$ (blue) and $\widehat{\beta}^{++}_{NT}$ (orange). Right panel: $\widehat{\delta}_{NT}$ (blue) and $\widehat{\delta}^{++}_{NT}$ (orange). Dashed, solid, dotted, dash-dotted lines correspond to DGPs (i)-(iv), respectively.

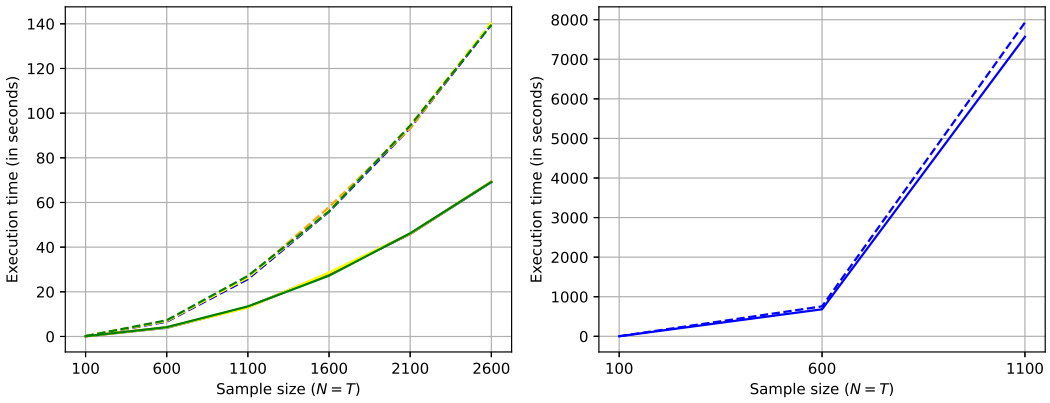
Reduced execution time by FPMLE and FPMLE^{++}. Figure 2.3 summarizes the execution time of our Python's implementation (`nlmfe`) of FPMLE^{++} (left panel) and STATA's `logitfe` (right panel). The statistics in both figures are computed using 10 Monte Carlo replications and with $N = T$. Given the similar precision of FPMLE and FPMLE^{++} after a relatively small number of iterations, we focus on FPMLE^{++} in this section and set the number of iterations to be 20 (and jointly with the usual numerical stopping criteria).[25] Overall, FPMLE^{++} largely outperforms `logitfe` in terms of execution time for all sizes of data sets. First, whether or not implementing post-estimation bias correction (dotted lines in Figure 2.3), FPMLE^{++} only uses a fractional execution time of that by `logitfe`. For instance, when $N = T = 1100$, FPMLE^{++} gives point estimates in less than 20s, while `logitfe` uses more than two hours. Second, for very large data sets (e.g., $N = T = 2600$), `logitfe` can not even run; in contrast, FPMLE^{++} still produces point estimates in about 1min. This remarkable time efficiency remains valid when $N \gg T$. See Table 2.7 in Section 2.7.11 for comparisons when $N = 5000$ and $T = 30$.[26]

Bootstrap inference procedure. We simulate a Poisson model with heterogeneous slopes and implement a split-sample jackknife bootstrap procedure in the spirit of Dhaene and Jochmans (2015) for the distributional features of β_{0i} as well as the

[25] In the Monte Carlo experiments and empirical illustrations, we jointly use a stopping criterion based on the variation of the objective function generated by the previous iterate (e.g., the iteration stops as soon as this variation is less than 10^{-5}). For FPMLE^{++}, we use a step size of $\nu^{(s)} \approx 1/(NT)$, or an Hessian step.

[26] In the same table, we also report the execution time of FPMLE. We find that FPMLE^{++} further reduces execution time relative to FPMLE.

FIGURE 2.3: Execution Time (in seconds) of FPMLE^{++} and `logitfe`, $N = T$



Notes: Left panel: Python's `nlmfe` implementation of FPMLE^{++} with $\nu = 1/(NT)$ and 20 iterations. Right panel: STATA's `logitfe`. DGPs (i)-(iv) in blue/orange/yellow/green, respectively. Solid lines: time to compute the estimates. Dashed lines: time to compute the Jackknife bias-corrected estimates. Elapsed time is compuputed using STATA's `timeit` and Python's `time.perf_counter()` commands on the ENSAE IP Paris's cluster (Intel(R) Xeon(R) Gold 6130 CPU, 2.10GHz, 256Gb RAM).

APEs.[27] The details of the Poisson model is in Appendix 2.7.4 and the implementation details of the Bootstrap procedure are in Section 2.7.11.

Table 2.1 summarizes the coverages of the percentile bootstrap jackknife confidence intervals (CIs) for the mean of β_{0i}, its standard deviation, and the APEs. Overall, the proposed bootstrap procedure achieves reasonable coverage. When the DGP satisfies Assumption 2.3.1 (DGPs (i) and (ii)), the coverages of the CIs for $\mathbb{E}(\beta_{0i})$ and the APE attain the desired levels. When the DGP violates the identification assumption (DGPs (iii) and (iv)), the coverages decrease relative to those in DGPs (i) and (ii), but are still reasonably large. We find that the coverages of the CIs for $\sqrt{\mathbb{V}(\beta_{0i})}$ are lower than the desired levels. They can be ameliorated by using asymmetric quantiles (e.g., using 4% and 99% quantiles to construct the CI with level 95%). See Table 2.3 for details.

Suggestions to practitioners. Due to the fast convergence and high precision of FPMLE/FPMLE^{++}, both algorithms provide good approximations of the MLE estimator and are useful for applied research in various settings. When the problem size is small or moderate and the concentrated MLE is still feasible, one can use FPMLE/FPMLE^{++} to accelerate the implementation of the MLE estimator. For

[27] In practice, when T or N is moderate, which is the case of our second empirical illustration, the splitted sample may be too small, leading to potentially unstable numerical performance. We test the performance of a straightforward percentile bootstrap procedure without the jackknife correction that makes use of the full sample and therefore minimizes the numerical instability. As expected, for DGPs (i) and (ii) that satisfy Assumption 2.3.1, the corresponding coverages are dominated by the procedure with the jacknife correction. In contrast, for DGPs (iii) and (iv) that violate Assumption 2.3.1, we do not find such dominance. An alternative method is analytical bias correction, which we leave for future research.

TABLE 2.1: Inference – Poisson Model with Heterogeneous Slopes

	$\widehat{\mathbb{E}}(\beta_{0i})$			$\sqrt{\widehat{\mathbb{V}}(\beta_{0i})}$			$\widehat{\text{APE}}$		
Coverage	.90	.95	.99	.90	.95	.99	.90	.95	.99
DGP									
i.	.9170	.9600	.9900	.5870	.6910	.8410	.9630	.9780	.9860
ii	.9450	.9820	.9950	.4980	.6370	.8080	.9800	.9920	.9990
iii.	.8040	.8560	.8900	.7810	.8310	.8840	.5530	.6790	.7960
iv.	.8250	.8850	.9380	.6430	.7210	.8420	.5970	.6970	.8330

Notes: Data are generated from the Poisson model described in Appendix 2.7.4 with $N = T = 50$. The coverages are computed based on 1,000 replications. For each repetition, we implement percentile Bootstrap jackknife CI's based on 200 Bootstrap samples. All computations are performed with FPMLE^{++} with at most 2 Hessian step iterations.

commonly used nonlinear models (e.g., logit, probit, Poisson), Assumption 2.3.2 is satisfied and the likelihood function has a unique global maximum. Then, both algorithms converge to the desired solution. When Assumption 2.3.2 may not hold, the likelihood function could have multiple local maxima. One can use both algorithms to fast back out the set of stationary points of the likelihood function (Proposition 2.7.6 in Section 2.7.9) and find out the global maximum, solving the practical challenge of multiple local maxima in the MLE approach (see the discussion after Theorem 2.3.2). When the problem size is large (or the computational resources are limited), running the concentrated MLE can be costly (e.g., $N = T \geq 1100$ in Figure 2.3). We suggest using either FPMLE or FPMLE^{++}. One can start with a small number of iterations (say, 20) and double check their numerical convergences by increasing the number of iterations (jointly with the usual numerical stopping criteria). Finally, if the problem is very large (e.g., $N = T = 2600$), it is possible that the implementation of FPMLE would become costly. In these cases, we suggest using FPMLE^{++}.

2.5 Empirical Illustrations

We demonstrate the empirical relevance of our proposed method by revisiting two classic studies: the determinants of trade flows (Helpman et al., 2008) and the causal relationship between institutional ownership and innovation (Aghion et al., 2013). Contrasting to the models in the original studies that impose homogeneity in the slope parameter capturing the causal relationship, we allow such slope to be individual-specific. In both illustrations, we find significant dispersion in the slope, suggesting important heterogeneity in the strength of the causal relationship. Moreover, the residual dispersion after controlling for individual's observed characteristics does not disappear, suggesting non-negligible unobserved heterogeneity in the slope parameter. Imposing homogeneous slopes and ruling out the heterogeneity may miss out the complexity in the underlying mechanism(s).

2.5.1 The Determinants of Trade Linkages and Flows

Helpman et al. (2008) estimate trade flows and explicitly take into account firm selection into export markets. Their method features a first step that estimates the establishment of exportation from one country to another using a binary model. Because of this step, they can then control for the fraction of firms that export (consistently estimated from the first step) and the selection effect due to zero trade flows when estimating the gravity equation in the second step. In the empirical application, this first step is implemented as following (see their equation 12 on page 455):

$$\mathbb{P}(T_{ij}=1|\text{dist}_{ij},w_{ij},\zeta_i,\xi_j)=\Phi\left(-\gamma\text{dist}_{ij}+w_{ij}'\kappa+\zeta_i+\xi_j\right),i,j=1,\ldots,N,i\neq j, \tag{2.5.1}$$

where $T_{ij}=1$ when country j exports to i and zero otherwise, dist_{ij} is the distance between i and j, w_{ij} is a vector of observed country-pair specific variables, ζ_i (ξ_j) is an importer (exporter) fixed effect, and Φ is the standard normal cumulative distribution function. According to their theoretical model, γ is interpreted as a constant elasticity of a firm's trade with respect to distance.

Different from the original setting, we allow γ to be country- and exporter-specific:

$$\mathbb{P}(T_{ij}=1|\text{dist}_{ij},w_{ij},\zeta_i,\xi_j)=\Phi(-\gamma_j^{\text{exp}}\text{dist}_{ij}+w_{ij}'\kappa+\zeta_i+\xi_j),i,j=1,\ldots,N,i\neq j. \tag{2.5.2}$$

Recent literature on international trade raises concerns about the assumption of constant trade elasticities that impose homogeneous effects of trade cost shifters (see Carrère et al. (2020); Chen and Novy (2021) for examples). The specification in (2.5.2) relaxes this assumption along two dimensions. First, it allows firms from different countries to react differently to the same change in trade cost shifters when exporting to the same third country. Second, two countries in a trade relationship, when exporting to the other, can react differently to the same change in the trade cost shifters that affects the trade in both directions. Furthermore, this specification is implied by a theoretical model along the lines of Helpman et al. (2008) with demand elasticity in the product market being country-specific.[28] We also consider another specification that allows γ to be country- and importer-specific:

$$\mathbb{P}(T_{ij}=1|\text{dist}_{ij},w_{ij},\zeta_i,\xi_j)=\Phi(-\gamma_i^{\text{imp}}\text{dist}_{ij}+w_{ij}'\kappa+\zeta_i+\xi_j),i,j=1,\ldots,N,i\neq j. \tag{2.5.3}$$

Similar to (2.5.2), the specification in (2.5.3) allows two countries in a trade relationship to react differently to the same change in the trade cost shifters that affects the trade in both directions. Moreover, (2.5.3) can also incorporate firm's heterogeneous reaction to the same change in trade cost shifters, depending on the country it exports to. In what follows, we estimate the first step of the method by Helpman et al.

[28] Concretely, denote by ε_j the demand elasticity in country j in their equation 2 on page 449. Then, the log of trade cost, $\ln\tau_{ij}$, enters the first (and the second) step with a coefficient ε_j-1. As a result, along the lines of their empirical specification, we can specify $(\varepsilon_j-1)\ln\tau_{ij}=\gamma_j^{\text{exp}}d_{ij}-u_{ij}$.

(2008) using (2.5.2) and (2.5.3), and quantify the extent to which the trade elasticity is heterogeneous among countries.[29]

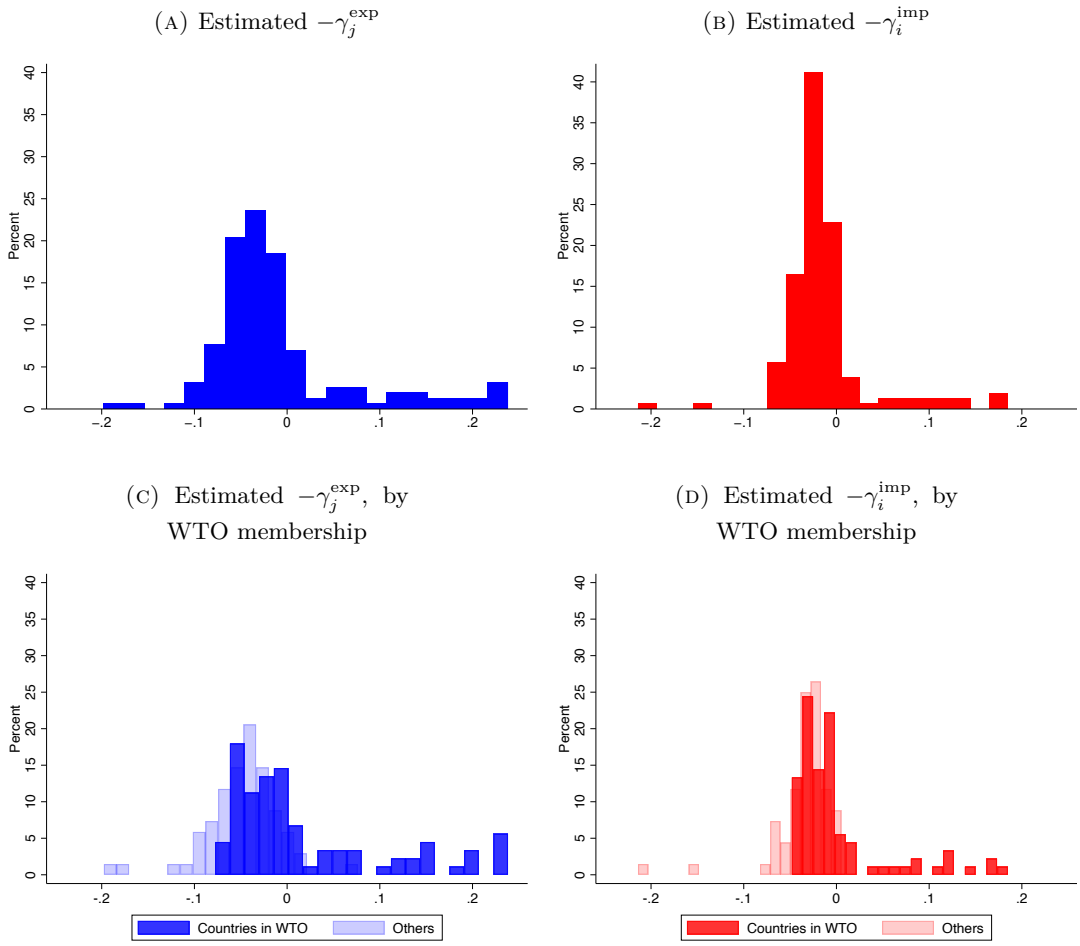
We estimate (2.5.2) and (2.5.3) using the 1986 worldwide trade data sample of Helpman et al. (2008) which include $N = 158$ countries. We remove Congo as an exporter from the sample because it did not export to anyone in 1986. This treatment leaves us with $24,649$ observations of trade flows (exportation) from country j to i.[30] We then obtain 157 estimated $\gamma_j^{\exp}$ and 158 estimated γ_i^{imp}. The average of $-\gamma_j^{\exp}$ (resp. $-\gamma_i^{\text{imp}}$) across countries is estimated to be -0.010 (resp. -0.014) and the corresponding marginal effects at the sample mean is -0.004 (resp. -0.006). We find non-negligible dispersion in $-\gamma_j^{\exp}$ and $-\gamma_i^{\text{imp}}$ (Figures 2.4(a) and (b)): the standard deviation of the former is estimated to be 0.078 and the latter is estimated to be 0.048, both of which are of greater magnitudes to the averages and statistically significant.[31] This dispersion can be partly explained by country characteristics that may determine the trade elasticity. In Figures 2.4(c) and (d), we plot the distribution of $-\gamma_j^{\exp}$ and $-\gamma_i^{\text{imp}}$ by country's WTO membership status. We find that as an exporting/importing country, j is less elastic with respect to distance if it is a member of the WTO (dark vs light blue/red in Figures 2.4(c) and (d). See also Table 2.4). The residual dispersion after controlling for such observed characteristics does not seem to disappear. We implement linear regressions of $-\gamma_j^{\exp}$ and $-\gamma_j^{\text{imp}}$ over country's WTO membership status and its geographic characteristics used in Helpman et al. (2008). The results are summarized in Table 2.4.

[29] Allowing for country-specific trade elasticity such as (2.5.2) and (2.5.3) may also change the second-step estimation in Helpman et al. (2008). First, and consistently, the trade elasticity parameter in the second stage will be also country-specific. Second, the estimated fraction of firms that export and inverse Mills ratio (if the first step is specified as a probit model), both of which are used as regressors in the second step, will take into account the heterogeneity in the estimated trade elasticity in the first step. Intuitively, the more significant heterogeneity in the trade elasticity is, the more the second step will be affected. Characterizing these consequences is beyond the scope of our methodology, which we leave for future research.

[30] We use the set of controls ϕ_{ij} in the second column of Table 1 of Helpman et al. (2008). Totally removing Congo from the sample does not significantly alter the results.

[31] The 95% symmetric percentile Bootstrap confidence intervals are $[0.065, 0.092]$ and $[0.036, 0.061]$, respectively.

FIGURE 2.4: Distribution of Estimated Trade Elasticity



Notes: Histograms based on 157 and 158 estimated $-\gamma_j^{\exp}$ and $-\gamma_i^{\text{imp}}$, respectively.

2.5.2 The Effects of Institutional Ownership on Innovation

Aghion et al. (2013) study how institutional ownership affects firm's innovation. The main empirical specification in the paper (Eq. 1 on page 280) is a two-way fixed effects Poisson count model:

$$\begin{aligned} &\text{CITES}_{it} \sim \text{Poisson}(\lambda_{it}), \\ &\lambda_{it} = \exp\left\{\beta \times \text{INSTIT}_{it} + \mathbf{x}_{it}'\alpha + \eta_i + \tau_t\right\}, \end{aligned} \tag{2.5.4}$$

where CITES_{it} is firm i's number of patents in period t weighted by future citations, INSTIT_{it} is the proportion of stock owned by institution investors, $\mathbf{x}_{it}$ is a vector of control variables (e.g., sales, firm size), η_i is firm-i specific fixed effect, and τ_t is period-t specific fixed effect. Slope β captures the causal effect of institutional ownership: 1 percentage point increase in INSTIT_{it} leads to firm i's number of patents in period t to change on average by 100β percentage points. Their main results (Table 1 on

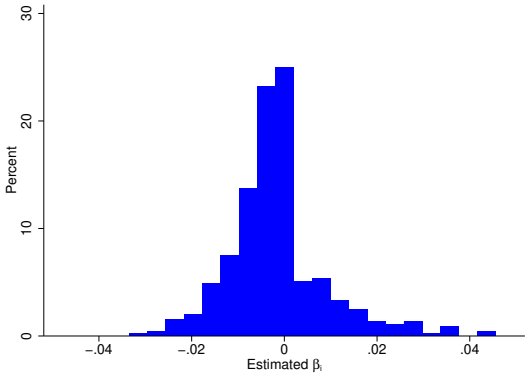
page 283) show that the coefficient of INSTIT_{it} is significantly positive, suggesting a positive impact of institution ownership on firm's innovation.

We allow the coefficient of INSTIT_{it} to be firm-i specific, i.e., β_i, in model (2.5.4).[32] This specification is plausible and coherent with the two micro-foundations in Aghion et al. (2013) (career concern and "lazy manager"): higher institution ownership will induce a higher probability of monitoring that incentives the manager to innovate more; however, because corporate structure may vary substantially across firms, the same change in the proportion of institution ownership may not produce the same amount of change in monitoring, leading to different incentives of innovation and therefore heterogeneous coefficient of institution ownership. This relaxation also raises interesting questions regarding the correlation between β_i and η_i, e.g., whether a more innovative firm (larger η_i) is more (or less) incentivized by the institutional monitoring to innovate (more positive β_i), and what are the drivers of the correlation (if there is any).

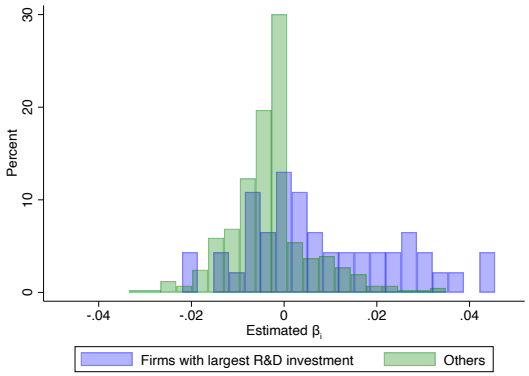
[32] We use the same set of controls x_{it} as column (5) of Table 1 in Aghion et al. (2013).

FIGURE 2.5: Distributions of $\widehat{\beta}_i$

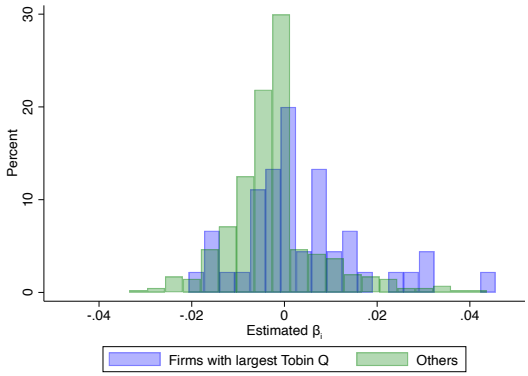
(A) Unconditional



(B) Conditional on R&D investment



(C) Conditional on Tobin's Q



Notes: Histograms based on 452 estimated β_i. Firms with largest R&D investment (Tobin'Q) are defined as those in the top 10% quantile of average R&D investment (Tobin' Q) over the time period in the data.

Figure 2.5 summarizes the unconditional distribution of $\widehat{\beta}_i$ (panel a) and some conditional distributions (panels b and c).[33] The estimated average of β_i is very close to zero (-0.0017). In addition, firms with greater R&D investment (panel b) and Tobin'Q (pabel c) are estimated to react more positively to institutional ownership than other firms. Importantly, we find a significant dispersion in $\widehat{\beta}_i$. We estimate $\widehat{\sigma}_\beta$ to be 0.0105, which is of the same magnitude as the effect found by Aghion et al. (2013).[34] We decompose $\widehat{\beta}_i$ over a set of firm i's characteristics, denoted by z_i, as follows:

$$\widehat{\beta}_i = z_i\gamma^\beta + \zeta_i^\beta. \tag{2.5.5}$$

[33]Different from the first empirical illustration, the panel data in the second one are unbalanced and the number of periods (9 years) is moderate compared to the number of firms (803 firms). As a result, the estimates could be noisier. Augmenting the data with observations from more periods will mitigate such noise in finite sample. Besides, the economic points of this illustration are still valid.

[34]In their Table 1, estimated β ranges from 0.005 to 0.01.

We can then estimate the extent to which the dispersion in $\widehat{\beta}_i$ is explained by the observed z_i and the unobserved component, ζ_i^β.[35] We find that ζ_i^β still explains around 62% of the total variation in $\widehat{\beta}_i$. See Table 2.5 for details.
Next, we assess the correlation between estimated β_i and η_i and find significantly positive correlation (0.231, with symmetric 95% percentile Bootstrap confidence interval being $[0.172, 0.806]$). To shed light on the drivers of this positive correlation, we decompose $\widehat{\eta}_i$ over the same set of firm i's characteristics z_i in (2.5.5):

$$\widehat{\eta}_i = z_i\gamma^\eta + \zeta_i^\eta. \tag{2.5.6}$$

By using both equations, we quantify the extent to which the correlation between $\widehat{\beta}_i$ and $\widehat{\eta}_i$ is driven by the observed correlation captured by $(z_i\gamma^\eta, z_i\gamma^\beta)$ and the unobserved correlation by $(\zeta_i^\beta, \zeta_i^\eta)$. We find that the correlation between $\widehat{\beta}_i$ and $\widehat{\eta}_i$ is not fully driven by the observed characteristics; the co-variance between $z_i\gamma^\eta$ and $z_i\gamma^\beta$ accounts for 49% of that between $\widehat{\beta}_i$ and $\widehat{\eta}_i$, and the correlation between ζ_i^β and ζ_i^η accounts for 51%.[36]

2.6 Conclusion

We study a class of nonlinear two-way fixed effects panel models that features individual-specific slopes in addition to the usual individual-specific and time-specific intercepts, and flexibly specified link function. The former is relevant when the researcher is interested in the distributional causal effects of covariates and their policy implications. The latter mitigates potential misspecification errors due to restrictions imposed on link function in empirical research. When both N and T are large, we prove that the fixed effects parameters and the link function can be nonparametrically identified using the strategy of compensating variable. We propose a novel iterative Gauss-Seidel estimation procedure that largely alleviates the challenge of dimensionality in the number of fixed effect parameters in the routine implementation of the MLE. We show that the procedure is numerically equivalent to the MLE under standard conditions. Extensive Monte Carlo simulations suggest its fast convergence and robust finite-sample performance in inference. We revisit two classic empirical studies in international trade (Helpman et al., 2008) and innovation (Aghion et al., 2013) to illustrate the empirical relevance of our method. Specifically, we investigate the extent to which the causal effect of interest is heterogeneous across individuals by allowing for (unobserved) heterogeneous slope parameters across countries/firms. We find non-negligible (unobserved) dispersion in trade elasticity and the effect of institutional ownership on firm innovation, respectively. These exercises emphasize

[35] We include the average of sales, R&D expenditure, Tobin's Q across time period, and sector dummies in z_i.
[36] See Tables 2.5 and 2.6 for more details of the decomposition.

the usefulness of the proposed method in capturing flexible (and unobserved) heterogeneity in the causal relationship of interest which may have important implications for the subsequent policy analysis.

2.7 Appendix

Notation: For any $p \geq 1$ and any two vectors x and y in $\mathbb{R}^p$, we let $\langle x, y \rangle$ denote the usual Euclidean inner product of x with y. Thus, the Euclidean norm is given by $\|x\| = \sqrt{\langle x, x \rangle}$. For any twice continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$, we let $\nabla f(x)$ (resp. $\nabla^2 f(x)$) denotes its gradient (resp. Hessian) at $x \in \mathbb{R}^p$. For a matrix A, we denote A' as the transpose of A. For a real symetric matrix $A \in \mathbb{R}^{n \times n}$, we let $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ denote its real eigenvalues. For any real matrix $A \in \mathbb{R}^{n \times m}$, $\|A\|_2 := \sqrt{\lambda_1(A'A)}$ denotes the spectral norm (i.e., the operator norm induced by the Euclidean norm), $\|A\|_{\mathrm{F}} := \sqrt{\mathrm{tr} A'A}$ denotes the Frobenius norm, and $\|A\|_{\max} := \max_{i=1,\ldots,n;j=1,\ldots,m} |A_{ij}|$ denotes the max norm.

2.7.1 Proof of Theorem 2.3.1

First, we focus on the observations corresponding to individual i and identify individual-fixed effects $\beta_i^{(1)}$, α_i, and $\beta_i^{(2)} - \beta_1^{(2)}$. For $\overline{y} \in \mathcal{Y}$ verifying Assumption 2.3.1(i), and $x \in \mathcal{X}_i$, define the following quantity:

$$\Gamma_i(\overline{y}; x) := \mathbb{E}\left[\mathbf{1}\{y_{it} = \overline{y}\} | x_{it} = x, \alpha_i, \beta_i\right]. \tag{2.7.1}$$

Then, under Assumption 2.3.1(ii), we can identify $\Gamma_i(\overline{y}; x)$ for each $x \in \mathcal{X}_i$. When $\mathcal{X}_i$ is discrete, $\Gamma_i(\overline{y}; x)$ is obtained by using the law of large numbers (LLN) and Slutsky's lemma.[37] When x_{it} has continuous components, it can obtained by using Nadaraya and Watson's estimator under standard regularity conditions on the density of x_{it} (see, e.g. Hansen, 2008). Using Assumption 2.3.1(iii), we obtain:[38]

$$\begin{aligned}
\Gamma_i(\overline{y}; x) &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}\{y_{it} = \overline{y}\} | x_{i1}, \ldots, x_{it-1}, x_{it} = x, \alpha_i, \beta_i, \xi_t\right] | x_{it} = x, \alpha_i, \beta_i\right] \\
&= \mathbb{E}\left[g(\overline{y}; \alpha_i + \xi_t + x'\beta_i) | x_{it} = x, \alpha_i, \beta_i\right] \\
&= \int g(\overline{y}; \alpha_i + \xi + x'\beta_i) dF_\xi(\xi; x^{(2)}).
\end{aligned}$$

Similarly,

$$\Gamma_1(\overline{y}; x) = \int g(\overline{y}; \alpha_1 + \xi + x'\beta_1) dF_\xi(\xi; x^{(2)})$$

[37] Concretely, one would rely on Bernstein's LLN over t's to identify $\Gamma_i(\overline{y}; x)$: given x, $\Gamma_i(\overline{y}; x)$ is obtained by aggregating $\mathbf{1}\{y_{it} = \overline{y}\}$ across countably many time periods as long as the correlation between $\mathbf{1}\{y_{it} = \overline{y}\}\mathbf{1}\{x_{it} = x\}$ and $\mathbf{1}\{y_{it'} = \overline{y}\}\mathbf{1}\{x_{it'} = x\}$ decreases to zero when $|t - t'| \to \infty$.

[38] The construction of $\Gamma_i(\cdot)$ does not require x_{it} to be different for individuals in the same time period t. Instead, for a given individual i, it aggregates the outcomes corresponding to this individual and with the same values of covariates $x_{it} = x$ (or, $x_t = x$ when $x_{it} = x_t$) across different periods.

is identified for $x \in \mathcal{X}_1$. Assumption 2.3.1(iv) ensures that we can find $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$ such that $(z_i(x^{(1)k}; x^{(2)k}), x^{(2)k}) \in \mathcal{X}_1$ with $k = 1, 2, 3$ and the matrix

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(1)1} \\ 1 & x^{(1)2} & x^{(1)2} \\ 1 & x^{(1)3} & x^{(1)3} \end{bmatrix}$$

is nonsingular. Fixing such $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$, we can then identify $(\tilde{x}^{(1)k}, x^{(2)k}) \in \mathcal{X}_1$ such that

$$\Gamma_1(\overline{y}; \tilde{x}^{(1)k}, x^{(2)k}) = \Gamma_i(\overline{y}; x^{(1)k}, x^{(2)k}).$$

Given $x^{(2)k}$, because of the strict monotonicity in Assumption 2.3.1(i) and the normalization $\beta_1^{(1)} = 1$, such $\tilde{x}^{(1)k}$ is unique. Moreover, by the definition of $z_i(x^{(1)}; x^{(2)})$, we know that

$$\Gamma_1(\overline{y}; z_i(x^{(1)k}; x^{(2)k}), x^{(2)k}) = \Gamma_i(\overline{y}; x^{(1)k}, x^{(2)k}).$$

As a result, we identify $z_i(x^{(1)k}; x^{(2)k}) = \tilde{x}^{(1)k}$ for $k = 1, 2, 3$, i.e.,

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(1)1} \\ 1 & x^{(1)2} & x^{(1)2} \\ 1 & x^{(1)3} & x^{(1)3} \end{bmatrix} \begin{pmatrix} \alpha_i \\ \beta_i^{(1)} \\ \beta_i^{(2)} - \beta_1^{(2)} \end{pmatrix} = \begin{pmatrix} \tilde{x}^{(1)1} \\ \tilde{x}^{(1)2} \\ \tilde{x}^{(1)3} \end{pmatrix}.$$

Solving this linear system identifies α_i, $\beta_i^{(1)}$ and $\beta_i^{(2)} - \beta_1^{(2)}$.

Second, we identify $\beta_i^{(2)}$ and ξ_t by further using Assumptions 2.3.1(ii) and (v). Note that $z_i(x^{(1)}; x^{(2)})$ is already identified for any $(x^{(1)}, x^{(2)}) \in \mathcal{X}_i$ and $i \in \mathbf{N}$. Fixing $t \in \mathbf{T}$ (and therefore conditional on (ξ_t, β_1)), because of the independence of $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i \in \mathbf{N}}$ in Assumption 2.3.1(ii), $\left\{(y_{it}, z_i(x_{it}^{(1)}; x_{it}^{(2)}), x_{it}^{(2)})\right\}_{i \in \mathbf{N}}$ are also independent. Similarly to (2.7.1), we then identify the following quantity:

$$\Gamma^t(y; z, x^{(2)}) := \mathbb{E}\left[\mathbf{1}\{y_{it} = \overline{y}\} | z_{it} = z, x_{it}^{(2)} = x^{(2)}, \xi_t, \beta_1^{(2)}\right].$$

for $(z, x^{(2)}) \in \mathcal{Z}_t$. Using (2.3.1), we have $\Gamma^t(y; z, x^{(2)}) = g(y; z + \beta_1^{(2)} x^{(2)} + \xi_t)$. Then, $g(y; z + \beta_1^{(2)} x^{(2)} + \xi_t)$ is identified for any $y \in \mathcal{Y}$, $(z, x^{(2)}) \in \mathcal{Z}_t$, and $t \in \mathbf{T}$.

Remark 2.7.1 *Suppose that $g(y; v)$ is real analytic with respect to $v \in \mathbb{R}$ for any $y \in \mathcal{Y}$ and $\mathcal{Z}_t$ contains an open set. Because $g(y; z + \beta_1^{(2)} x^{(2)} + \xi_t)$ is identified for $(z, x^{(2)}) \in \mathcal{Z}_t$, $g(y; \alpha_i + x'\beta_i^{(1)} + \xi_t)$ is then identified for $x \in \mathcal{X}_i$ such that $z_i(x^{(1)}; x^{(2)}) \in \mathcal{Z}_t$. Note that $z_i(\cdot)$ is continuous in x and $\mathcal{Z}_t$ has an open subset. Then, the preimage of $\mathcal{Z}_t$ also contains an open subset and $g(y; \alpha_i + x'\beta_i + \xi_t)$ is identified in this open subset of $\mathcal{X}_i$. Due to the unique continuation of the real analytic function and the real analyticity of $g(y; \alpha_i + x'\beta_i + \xi_t)$ with respect to x, we can then identify $g(y; \alpha_i + x'\beta_i + \xi_t)$ as well as its derivatives with respect to x for $x \in \mathbb{R}^2$.*[39]

[39] The real analyticity of $g(y; \alpha_i + x'\beta_i + \xi_t)$ with respect to x is a result of that of $g(y; v)$ with respect to v and $v = \alpha_i + x'\beta_i + \xi_t$ with respect to x.

Because of Assumption 2.3.1(v)a, we can find $t \in \mathbf{T}$, $r \in \mathbb{R}$, and two $(z, x^{(2)}), (z', x^{(2')}) \in \mathcal{Z}_t$, such that $x^{(2)} \neq x^{(2')}$ and $g(y; r+\xi_t) = g(y; z+\beta_1^{(2)}x^{(2)}+\xi_t) = g(y; z' + \beta_1^{(2)}x^{(2')} + \xi_t)$. Using Assumption 2.3.1(i) and setting $y = \bar{y}$, we obtain $z + \beta_1^{(2)}x^{(2)} + \xi_t = z' + \beta_1^{(2)}x^{(2')} + \xi_t$, identifying $\beta_1^{(2)}$ (and therefore $\beta_i^{(2)}$). This further identifies $g(y; z + \beta_1^{(2)}x^{(2)} + \xi_t)$ for any $(z, x^{(2)})$ as long as $\mathrm{P}_{(1,\beta_1^{(2)})}(z, x^{(2)}) \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it})$ for all $t \in \mathbf{T}$ (and $g(y; v + \xi_t)$ for any $v \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it})$ for all $t \in \mathbf{T}$).

Assumption 2.3.1(v)b ensures for any $t \in \mathbf{T}$,

$$\left\{g(\bar{y}; z + \beta_1^{(2)}x^{(2)} + \xi_t) : \mathrm{P}_{(1,\beta_1^{(2)})}(z, x^{(2)}) \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it})\right\}$$
$$\cap$$
$$\left\{g(\bar{y}; z + \beta_1^{(2)}x^{(2)}) : \mathrm{P}_{(1,\beta_1^{(2)})}(z, x^{(2)}) \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{i1})\right\} \neq \emptyset.$$

Then, using Assumption 2.3.1(i), we can then find $(z, x^{(2)})$ with $\mathrm{P}_{(1,\beta_1^{(2)})}(z, x^{(2)}) \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it})$, and $(z', x^{(2')})$ with $\mathrm{P}_{(1,\beta_1^{(2)})}(z', x^{(2)'}) \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{i1})$, such that

$$z + \beta_1^{(2)}x^{(2)} + \xi_t = z' + \beta_1^{(2)}x^{(2')}.$$

We then identify ξ_t by $z' - z + \beta_1^{(2)}(x^{(2')} - x^{(2)})$.

Finally, the identification of ξ_t and $g(y; v + \xi_t)$ for any $v \in \cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it})$ allow to identify $g(y; v)$ as a function of $\mathcal{Y} \times \left(\cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it}) + \xi_t\right)$. We then identify $g(y; v)$ in $\mathcal{Y} \times \cup_{t\in\mathbf{T}}\left(\cap_{i\in\mathbf{N}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it}) + \xi_t\right)$.

Pairwise Compensation

Consider the following generalization of Definition 2.3.1.

Definition 2.7.1 (Pairwise compensable) *Individual i is compensable at $(x^{(1)}, x^{(2)}) \in \mathcal{X}_i = Supp(x_{it}|\alpha_i, \beta_i)$ by individual i' if $(z_i(x^{(1)}; x^{(2)}) \in \mathcal{X}_{i'}$.*

Suppose that there exists a sequence $(i_1, i_2, \ldots)$ with $\{i_1, i_2, \ldots\} = \mathbf{N}$ such that i_n is compensable by i_{n+1} at least at three points in $\mathcal{X}_{i_n}$ with the same rank condition in Assumption 2.3.1(iv). Then, using the same identification argument in the proof of Theorem 2.3.1, we can identify $\alpha_{i_n} - \alpha_{i_{n+1}}$, $\beta_{i_n}^{(1)} - \beta_{i_{n+1}}^{(1)}$, and $\beta_{i_n}^{(2)} - \beta_{i_{n+1}}^{(2)}$ for all $n = 1, 2, \ldots$. Without loss of generality, suppose individual i is indexed by i in this sequence for any i. Then, we can identify α_i, $\beta_i^{(1)}$, and $\beta_i^{(2)} - \beta_1^{(2)}$ by $\sum_{r=1}^{i-1}(\alpha_{r+1} - \alpha_r)$, $\sum_{r=1}^{i-1}\left(\beta_{r+1}^{(1)} - \beta_r^{(1)}\right)$, and $\sum_{r=1}^{i-1}\left(\beta_{r+1}^{(2)} - \beta_r^{(2)}\right)$, respectively.

The key insight behind pairwise compensation is that one can "order" all individuals in a way that one individual can be compensated by the next one in the sequence. This could hold if the parameters of interest have a monotonic dependence on individuals' characteristics, i.e., shape restriction. We illustrate this point in the

following example and show how such restriction can attenuate support requirement on $x^{(1)}$.

Example 6 (Support condition in pairwise compensation) *We use the same setting as in Example 5 but drop $x^{(2)}$ from the model. Suppose that $\beta_i^{(1)}$, the parameter of disutility of price, is continuous with respect to w_i, individual i's income, denoted by $\beta(w_i)$, and decreases with w_i, i.e., a richer individual is less sensitive to price change. Moreover, α_i depends continuously on w_i, denoted by $\alpha(w_i)$.*

Start with the individual with the highest income $\max\{w_i\}$ whose $\beta(\max\{w_i\})$ is then the smallest. Now consider the individual whose income is slightly below, say $\max\{w_i\} - \epsilon$. Then, this individual's $\beta(\max\{w_i\} - \epsilon)$ is slightly greater than $\beta(\max\{w_i\})$ and the corresponding $\alpha(\max\{w_i\} - \epsilon)$ is slightly different from $\alpha(\max\{w_i\})$. Then, the individual with the highest income can be compensated by the one with the slightly lower income if there exists $x^{(1)} \in (a, A)$,

$$\frac{\alpha(\max\{w_i\}) - \alpha(\max\{w_i\} - \epsilon)}{\beta(\max\{w_i\} - \epsilon)} + \frac{\beta(\max\{w_i\})}{\beta(\max\{w_i\} - \epsilon)} x^{(1)} \in (a, A). \tag{2.7.2}$$

Because $\frac{\alpha(\max\{w_i\}) - \alpha(\max\{w_i\} - \epsilon)}{\beta(\max\{w_i\} - \epsilon)} \approx 0$ and $\frac{\beta(\max\{w_i\})}{\beta(\max\{w_i\} - \epsilon)} \approx 1$, the compensating variable $\frac{\alpha(\max\{w_i\}) - \alpha(\max\{w_i\} - \epsilon)}{\beta(\max\{w_i\} - \epsilon)} + \frac{\beta(\max\{w_i\})}{\beta(\max\{w_i\} - \epsilon)} x^{(1)}$ is always in a neighborhood of $x^{(1)}$. Consequently, as long as $A > a$, (2.7.2) always holds.

We can repeat this argument to another individual with a slightly lower w_i than $\max\{w_i\} - \epsilon$ and show that she is compensable by the individual with income $\max\{w_i\} - \epsilon$, forming the required sequence of compensation. Note that we only require $A > a$ and the size of the support $A - a$ can be arbitrarily small, which differs from the support condition in Assumption 2.3.1(iv).

2.7.2 Proof of Theorem 2.3.2

Preliminary results

We first recall classical results from the optimization literature. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and $X \subset \mathbb{R}^n$. f is said (μ, X)-strongly convex if there exists a constant $\mu > 0$ such that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y \in X. \tag{2.7.3}$$

f is said (L, X)-smooth if there exists a constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in X. \tag{2.7.4}$$

Lemma 2.7.1

1. *f is (μ, X)-strongly convex if and only if*

$$\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \mu \|x - y\|^2, \quad \forall x, y \in X. \tag{2.7.5}$$

2. *Suppose that f is twicely differentiable on $\mathbb{R}^n$.*

   (a) *$\nabla^2 f(x) \gtrsim \mu I$ for some $\mu > 0$ and all $x \in X$ if and only if f is (μ, X)-strongly convex.*

   (b) *If $\nabla^2 f(x) \lesssim LI$ for some $L > 0$ and all $x \in X$, then f is (L, X)-smooth.*

The proof of Lemma 1 can be found in Section 2.7.10.

Proposition 2.7.2 (Bertsekas (2016), Proposition 3.7.1) *Consider the problem*

$$\begin{aligned} &\min \ f(x) \\ &\text{subject to } x \in X, \end{aligned}$$

where X is a Cartesian product $X = X_1 \times \cdots \times X_m$ of closed convex subsets $X_i \subset \mathbb{R}^{n_i}$ such that $\sum_{i=1}^m n_i = n$. Suppose that for each $x = (x_1, \ldots, x_m) \in X$ and $i \in \{1, \ldots, m\}$, $y \mapsto f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_m)$ attains a unique minimum $\overline{y}$ over X_i, and is monotonically nonincreasing in the interval from x_i to $\overline{y}$. Let $\left\{x^k\right\}$ be the sequence generated by the block coordinate descent method which generates the next iterates $x^{k+1} = (x_1^{k+1}, \ldots, x_m^{k+1})$, given the current iterate $x^k = (x_1^k, \ldots, x_m^k)$, according to the iteration

$$x_i^{k+1} \in \underset{y \in X_i}{\arg\min} \ f(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, y, x_{i+1}^k, \ldots, x_m^k), \quad i = 1, \ldots, m.$$

Then, every limit point x^* of $\left\{x^k\right\}$ is a stationary point, i.e.,

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in X. \tag{2.7.6}$$

Proof of Theorem 2.3.2: FPMLE

We proceed in two steps. In a first step, we apply Proposition 2.7.2 to $f = -\mathscr{L}_{NT}$ to show that any limit point of the sequence of iterates generated by FPMLE, $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,\ldots}$, is a stationary point of $-\mathscr{L}_{NT}$. In a second step, we show that such a limit point exists and that $\widehat{\theta}_{NT}^{\text{MLE}}$ is the unique stationary point of $-\mathscr{L}_{NT}$.

Step 1: any limit point of $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,\ldots}$ is a stationary point of $-\mathscr{L}_{NT}$. We show that the conditions of Proposition 2.7.2 hold for $f = -\mathscr{L}_{NT}$, $m = N + T$, $n_1 = \cdots = n_{N+T-1} = 1, n_{N+T} = K, n = K + N + T - 1$, $X_1 \times \cdots \times X_{N-1} = \mathcal{A}_2 \times \cdots \times \mathcal{A}_N$, $X_N \times \cdots \times X_{N+T-1} = \Xi_1 \times \cdots \times \Xi_T$, and $X_{N+T} = \mathcal{B}$. By Assumption 2.3.2(i), $X = X_1 \times \cdots \times X_{N+T} = \Theta_{NT}$ is a Cartesian product of closed convex sets. By Assumption 2.3.2(ii), f is continuously differentiable over X. Let $(\alpha, \xi, \beta) \in \Theta_{NT}$,

and define the $m = N + T$ real-valued functions

$$\begin{array}{lll} f_{i,\alpha_{-(i+1)},\xi,\beta} & : a \in \mathcal{A}_{i+1} & \mapsto -\mathscr{L}_{NT}(\alpha_2, \ldots, \alpha_i, a, \alpha_{i+2}, \ldots, \alpha_N, \xi, \beta), \quad i = 1, \ldots, N-1, \\ f_{N+t-1,\alpha,\xi_{-t},\beta} & : e \in \Xi_t & \mapsto -\mathscr{L}_{NT}(\alpha, \xi_1, \ldots, \xi_{t-1}, e, \xi_{t+1}, \ldots, \xi_T, \beta), \quad t = 1, \ldots, T, \\ f_{N+T,\alpha,\xi} & : b \in \mathcal{B} & \mapsto -\mathscr{L}_{NT}(\alpha, \xi, b). \end{array}$$

The fact that each sets

$$\arg\min_{a \in \mathcal{A}_{i+1}} f_{i,\beta,\alpha_{-(i+1)},\xi}(a), \ \arg\min_{e \in \Xi_t} f_{N+t-1,\beta,\alpha,\xi_{-t}}(e), \ \arg\min_{b \in \mathcal{B}} f_{N+T,\alpha,\xi}(b) \tag{2.7.7}$$

are (nonempty) singletons follows from coercivity and strict concavity of each function $f_{i,\beta,\alpha_{-(i+1)},\xi}$, $f_{N+t-1,\beta,\alpha,\xi_{-t}}$, and $f_{N+T,\alpha,\xi}$, and standard functional analysis arguments (see Section 2.7.6). Finally, the monotonicity condition required to apply Proposition 2.7.2 follows from the strict convexity of f.

Step 2: $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,\ldots}$ admits a limit point and $\widehat{\theta}_{NT}^{\text{MLE}}$ is the unique stationary point of $-\mathscr{L}_{NT}$. By coercivity of $-\mathscr{L}_{NT}$, the level sets of $-\mathscr{L}_{NT}$ are bounded (and hence compact). Hence, there exists at least one limit point to the sequence $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,\ldots}$. Because $-\mathscr{L}_{NT}$ is convex and differentiable over Θ_{NT}, the set of stationary points of $-\mathscr{L}_{NT}$ is (see, e.g., Theorem 1.1.3(a) in Bertsekas, 2016):

$$\Theta_{NT}^* := \left\{\theta \in \Theta_{NT} : \left\langle \nabla \mathscr{L}_{NT}(\theta), \widetilde{\theta} - \theta \right\rangle \leq 0, \ \forall \widetilde{\theta} \in \Theta_{NT}\right\}.$$

Again, by coercivity and strict convexity of $-\mathscr{L}_{NT}$, we have $\Theta_{NT}^* = \left\{\widehat{\theta}_{NT}^{\text{MLE}}\right\}$.

Proof of Theorem 2.3.2: FPMLE^{++}

To simplify the exposition, consider for the moment the case with an homogeneous slope coefficient β. We extend the proof to the case with heterogeneous coefficients $(\beta_i)_{i \in \mathbf{N}}$ at the end of this section.

For any $\theta = (\alpha, \xi, \beta) \in \Theta_{NT}$, we let $\theta_1 = \alpha_2, \ldots, \theta_{N-1} = \alpha_N, \theta_N = \xi_1, \ldots, \theta_{N+T-1} = \xi_T$, and $\theta_{N+T} = \beta$. Let X_i be the reference space of θ_i (e.g., $X_{N+T} = \mathcal{B}$) and $X = X_1 \times \cdots \times X_{N+T} = \Theta_{NT}$. Let $f = -\mathscr{L}_{NT}$ and $\nabla_i f(\theta), \nabla_i^2 f(\theta)$ denote the gradient and Hessian operators respectively applied to f restricted to the coordinates of bloc i for $i \in \{1, \ldots, N+T\}$. The proof consists in verifying that FPMLE^{++} meets the conditions of Theorem 3.1 in Luo and Tseng (1993), a high-level result establishing linear convergence rates for a large class of feasible descent algorithms applied to the problem of finding stationary points of a continuously differentiable function whose gradient is Lipschitz continuous.[40]

[40] Note that we do not apply Luo and Tseng (1993)'s Proposition 3.4 for FPMLE because they require more stringent conditions than Proposition 2.7.2. A recent general treatment of block-coordinate gradient descent algorithms similar to FPMLE^{++} is given in Beck and Tetruashvili (2013). We do not use their results because they assume $(\mu, \mathbb{R}^{N+T+K-1})$-strong convexity and $(L, \mathbb{R}^{N+T+K-1})$-smoothness of $\mathscr{L}_{NT}$ which rarely holds in our econometric examples, except for rare exceptions (e.g., $(L, \mathbb{R}^{N+T+K-1})$-smoothness of $\mathscr{L}_{NT}$ holds for the logit model).

First, Luo and Tseng (1993)'s Assumption A holds because f is convex. Second, by Assumptions 2.3.2(i) and 2.3.2(iii), $X = \Theta_{NT}$ is compact and convex as a Cartesian product of compact and convex sets, and the functions $\theta \mapsto \lambda_1(\nabla^2 f(\theta))$ and $\theta \mapsto \lambda_{N+T}(\nabla^2 f(\theta))$ are continuous and strictly positive on Θ_{NT}. By the extreme value theorem and Lemma 2.7.1.2, it follows that f is (μ, Θ_{NT})-strongly convex and (L, Θ_{NT})-smooth for some $\mu, L > 0$.[41] Theorem 3.1 in Pang (1987) (whose Assumption (B) holds by Lemma 2.7.1.1) ensures that Luo and Tseng (1993)'s Assumption B holds with $\tau = (L+1)/\mu$. By similar arguments, there exists a sequence of strictly positive constants $(\mu_i)_i$ such that, for any $\theta \in X$, any $i \in \{1, \dots, N+T\}$, and any $\theta_i' \in X_i$,

$$f(\theta_1, \dots, \theta_{i-1}, \theta_i', \theta_{i+1}, \dots, \theta_{N+T}) - f(\theta) + \langle \nabla_i f(\theta), \theta_i - \theta_i' \rangle \geq \mu_i \left\| \theta_i' - \theta_i \right\|^2, \quad (2.7.8)$$

i.e., condition C in Luo and Tseng (1993) holds with $\gamma = \bar{\mu} := \min_i \mu_i$. Third, it remains to show that equations (3.1)-(3.3) in Luo and Tseng (1993) hold. Fix any index s. By definition of FPMLE^{++} iterates, we have

$$\theta_i^{(s+1)} = \left[\theta_i^{(s)} - \nu^{(s)} \nabla_i f(\theta_1^{(s+1)}, \dots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \dots, \theta_{N+T}^{(s)}) \right]_{X_i}^+, \quad i = 1, \dots, N+T.$$

Since $X = X_1 \times \cdots \times X_{N+T}$ is a Cartesian product of boxes, we have

$$\theta^{(s+1)} = \left[\theta^{(s)} - \nu^{(s)} \nabla f(\theta^{(s)}) + e^{(s)} \right]_X^+, \quad (2.7.9)$$

where $e^{(s)}$ is the vector in $\mathbb{R}^{N+T}$ whose ith component is

$$e_i^{(s)} = \nu^{(s)} \left[\nabla_i f(\theta^{(s)}) - \nabla_i f(\theta_1^{(s+1)}, \dots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \dots, \theta_{N+T}^{(s)}) \right].$$

Under Assumptions 2.3.2(i) and 2.3.2(iii), there exist positive constants $L_{11}, \dots, L_{N+TN+T}$ such that, for all i, j,

$$\left\| \nabla_i f(x) - \nabla_i f(x_1, \dots, x_{j-1}, y_j, x_{j+1}, \dots, x_{N+T}) \right\| \leq L_{ij} \left\| y_j \right\|, \quad \forall (x, y_j) \in X \times X_j. \quad (2.7.10)$$

Let $\bar{\nu} := \sup_s \nu^{(s)} \leq 1/\bar{L} < +\infty$ where $\bar{L} := \max_{i,j} L_{ij}$. By the triangle inequality and the Lipschitz conditions (2.7.10), we have

$$\begin{aligned} \left\| e_i^{(s)} \right\| &\leq \nu^{(s)} \left\| \nabla_i f(\theta_1^{(s+1)}, \dots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \dots, \theta_{N+T-1}^{(s)}) - \nabla_i f(\theta^{(s)}) \right\| \\ &\leq \nu^{(s)} \sum_{j=1}^{i} L_{ij} \left\| \theta_j^{(s)} - \theta_j^{(s+1)} \right\| \\ &\leq i \bar{\nu} \bar{L} \left\| \theta^{(s)} - \theta^{(s+1)} \right\|, \end{aligned}$$

[41] These constants can be easily derived as functions of the data and Θ_{NT} in common settings (e.g., logit/probit/Poisson).

where the last equality uses the uniform bounds $\nu^{(s)} \leq \bar{\nu}$, $L_{ij} \leq \bar{L}$, and $\left\|\theta_j^{(s)} - \theta_j^{(s+1)}\right\| \leq \left\|\theta^{(s)} - \theta^{(s+1)}\right\|$. By the triangle inequality again, we obtain

$$\left\|e^{(s)}\right\| \leq \sum_{i=1}^{N+T} \left\|e_i^{(s)}\right\| \leq \bar{\nu}\bar{L} \left\|\theta^{(s)} - \theta^{(s+1)}\right\| \sum_{i=1}^{N+T} i \leq \frac{(N+T)(N+T+1)}{2} \left\|\theta^{(s)} - \theta^{(s+1)}\right\|. \tag{2.7.11}$$

Equations (2.7.9) and (2.7.11) show that (3.1)-(3.2) in Luo and Tseng (1993) hold with $\kappa_1 = \frac{(N+T)(N+T+1)}{2}$ and $\alpha^r = \nu^{(r)}$ for all r. We now show that (3.3) in Luo and Tseng (1993) holds. Let consider a fixed iteration s. It suffices to show that for all $i \in \{1, \ldots, N+T\}$,

$$\left\langle \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \leq 0. \tag{2.7.12}$$

Equation (3.3.) with $\kappa_2 = \bar{\mu}$ in Luo and Tseng (1993) then immediately follow by summing (2.7.8) over $i \in \{1, \ldots, N+T\}$ and cancelling terms:

$$f(\theta^{(s)}) - f(\theta^{(s+1)}) \geq \bar{\mu} \left\|\theta^{(s)} - \theta^{(s+1)}\right\|^2.$$

Let us show (2.7.12). For each $i \in \{1, \ldots, N+T\}$, Taylor-Lagrange formula with integral remainder ensures

$$\begin{aligned}
&\nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)}) \\
&\quad = \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)}) \\
&\qquad + \int_0^1 \nabla^2 f_i(\theta_i^{(s)} + t(\theta_i^{(s+1)} - \theta_i^{(s)}))(\theta_i^{(s+1)} - \theta_i^{(s)}) dt,
\end{aligned}$$

where we define $f_i : x \mapsto f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, x, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)})$. It follows that

$$\begin{aligned}
&\left\langle \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\
&\quad = \left\langle \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\
&\qquad + \left\langle \int_0^1 \nabla^2 f_i(\theta_i^{(s)} + t(\theta_i^{(s+1)} - \theta_i^{(s)}))(\theta_i^{(s+1)} - \theta_i^{(s)}) dt, \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle.
\end{aligned} \tag{2.7.13}$$

Eq. (2.7.10) with $j = i$ implies that f_i is (L_{ii}, X_i)-smooth so that by Lemma 2.7.1.2(b), we have

$$\nabla^2 f_i(x_i) \lesssim \bar{L} I, \quad \forall x_i \in X_i. \tag{2.7.14}$$

Since X_i is convex, $\theta_i^{(s)}+t(\theta_i^{(s+1)}-\theta_i^{(s)}) \in X_i$. Then, by linearity of the scalar product, (2.7.14), and monotonicity of the integral together, we obtain

$$\begin{aligned}
&\left\langle \int_0^1 \nabla^2 f_i(\theta_i^{(s)} + t(\theta_i^{(s+1)} - \theta_i^{(s)}))(\theta_i^{(s+1)} - \theta_i^{(s)})dt, \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\
&= \int_0^1 \left\langle \nabla^2 f_i(\theta_i^{(s)} + t(\theta_i^{(s+1)} - \theta_i^{(s)}))(\theta_i^{(s+1)} - \theta_i^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle dt \\
&\leq \int_0^1 \bar{L} \left\| \theta_i^{(s+1)} - \theta_i^{(s)} \right\|^2 dt \\
&= \bar{L} \left\| \theta_i^{(s+1)} - \theta_i^{(s)} \right\|^2 .
\end{aligned}$$

Plugging this result into (2.7.13) yields

$$\begin{aligned}
&\left\langle \nabla f_i(\theta_1^{(s+1)}, \ldots, \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\
&\leq \left\langle \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle + \bar{L} \left\| \theta_i^{(s+1)} - \theta_i^{(s)} \right\|^2 .
\end{aligned} \tag{2.7.15}$$

Notice that $\theta_i^{(s+1)} - \theta_i^{(s)} = -\nu^{(s)} \left[\nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right]_{X_i}^{+}$. Because X_i is compact with 0 in its interior, there exists $a \in (0, +\infty)$ such that $a\nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \in X_i$. By linearity of $[\cdot]_{X_i}^{+}$

$$\begin{aligned}
a(\theta_i^{(s+1)} - \theta_i^{(s)}) &= -a\nu^{(s)} \left[\nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right]_{X_i}^{+} \\
&= -\nu^{(s)} \left[a\nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right]_{X_i}^{+} \\
&= -a\nu^{(s)} \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}).
\end{aligned} \tag{2.7.16}$$

Hence, multiplying the first term of (2.7.15) by a gives

$$\begin{aligned}
&\left\langle \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}), a(\theta_i^{(s+1)} - \theta_i^{(s)}) \right\rangle \\
&= -a\nu^{(s)} \left\| \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right\|^2 .
\end{aligned} \tag{2.7.17}$$

Next, since orthogonal projection is a contraction, using (2.7.10) and multiplying the second term of (2.7.15) by a we obtain

$$\begin{aligned}
a \left\| \theta_i^{(s+1)} - \theta_i^{(s)} \right\|^2 &= a\bar{L}\nu^{(s)2} \left\| \left[\nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right]_{X_i}^{+} \right\|^2 \\
&\leq a\bar{L}\nu^{(s)2} \left\| \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right\|^2 .
\end{aligned} \tag{2.7.18}$$

Finally, multiplying (2.7.15) by a, combining (2.7.17)-(2.7.18), and dividing by a yields

$$\left\langle \nabla f_i(\theta_1^{(s+1)}, \ldots, \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, \ldots, \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \leq \nu^{(s)}(\nu^{(s)}\bar{L} - 1) \left\| \nabla_i f(\theta_1^{(s+1)}, \ldots, \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, \ldots, \theta_{N+T}^{(s)}) \right\|^2 .$$

Because $0 \leq \nu^{(s)} = \nu \leq 1/\bar{L}$ the right hand-side is negative, which proves (2.7.12).

Taken together, the above results show that the conditions of Theorem 3.1 in Luo

and Tseng (1993) are verified with $\kappa_1 = \frac{(N+T)(N+T+1)}{2}$, $\kappa_2 = \bar{\mu}$, and $\alpha^r = \nu^{(r)}$ for all r. Then, their Theorem implies that $\widehat{\theta}_{NT}^{(s)++}$ converge R-linearly to $\widehat{\theta}_{NT}^{\text{MLE}}$, i.e., there exist constants $C_2 > 0$ and $\gamma < 1$ such that

$$\left\|\widehat{\theta}_{NT}^{++(s)} - \widehat{\theta}_{NT}^{\text{PJML}}\right\| \leq C_2\gamma^s.$$

The proof of Theorem 2.3.2 is complete.

Numerical convergence in the presence of heterogeneous slopes. We need to first strengthen Assumption 2.3.2 to accommodate this case. Because the dimension of the parameters becomes $N(K+1)+T-1$, we then need to extend the requirements on the likelihood function, in particular, the strict concavity and coercivity in Assumption 2.3.2(ii) and the smoothness in Assumption 2.3.2(iii) to the new space of parameters $\mathbf{R}^{N(K+1)+T-1}$. Using these conditions, the difference from the proof in the case of homogeneous slopes lies in the definitions of $\mu, L, \bar{L}, \bar{\mu}$ and thus κ_1, κ_2 and in turn C_2 and γ. In effect, both objects should be defined on the basis of the number of parameters $N(K+1)+T-1$ and all the other arguments go through. For FPMLE^{++}, as a consequence of the increased number of parameters, the constant learning rate $\nu^{(s)} = \nu$ will become smaller to satisfy the requirement that it is no greater than $1/\bar{L}$.

2.7.3 Consistency in the Presence of Heterogeneous Slopes

Let $\boldsymbol{x} = \{x_{it} : (i,t)\}$, $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_N^0)'$, $\boldsymbol{\xi}^0 = (\xi_1^0, \ldots, \xi_T^0)'$, $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_N^0)' \in \mathbb{R}^{N\times K}$, $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}^0)$, $\pi_{it}^0 = \alpha_i^0 + \xi_t^0$, $z_{it}^0 = x_{it}'\beta^0 + \pi_{it}^0$, and $\partial_{z^q}\ell_{it} = \partial_{z^q}\ell_{it}(z_{it}^0)$. Let $\mathcal{Z} = \text{Supp}(z_{it}^0)$. The two-way fixed effects estimator $\widehat{\boldsymbol{\theta}}$ verifies

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta_{NT}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{NT}\left\{\sum_{i=1}^{N}\sum_{t=1}^{T}\ell_{it}(x_{it}'\beta_i + \pi_{it}) - Q_{NT}(\boldsymbol{\alpha}, \boldsymbol{\xi})\right\}, \tag{2.7.19}$$

where function $\ell_{it}(\cdot)$ encapsulates individual i's response in time t, y_{it} and $Q_{NT} > 0$ is any penalty function that enforces the chosen normalization of the fixed effects such that $Q_{NT}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\xi}}) = Q_{NT}(\boldsymbol{\alpha}^0, \boldsymbol{\xi}^0) = 0$.

Assumption 2.7.1

(i). (Model) y_{it} is distributed as

$$y_{it}|\boldsymbol{x}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}^0 \sim \exp[\ell_{it}(x_{it}'\beta_i^0 + \pi_{it}^0)],$$

and conditional on $(\boldsymbol{x}, \boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}^0)$, y_{it} is independent across (i,t).

(ii). (Compactness) For all N, T, Θ_{NT} is compact and $\boldsymbol{\theta}^0, \widehat{\boldsymbol{\theta}}$ lie in the interior of Θ_{NT}.

(iii). (Asymptotics) As N and T tend to infinity: $N/T \to \kappa^2 \in (0, +\infty)$.

(iv). (Smoothness, tails, and moments) $z \mapsto \ell_{it}(z)$ *is four times continuously differentiable over* $\mathcal{Z}$ *a.s. and there exist constants* $C_1, C_2, C_3 > 0$ *such that, for all* $k, i, t, N, T, q \leq 4$, $\max_{i,t} \mathbb{E}[|\partial_{z^q} \ell_{it}(z_{it}^0)|^{8+\nu}] \leq C_1$ *for some* $\nu > 0$, $\mathbb{E}[\exp(\lambda \left|\partial_z \ell_{it}(z_{it}^0) x_{it}^{(k)}\right|)] \leq \exp(\lambda C_2)$ *for all* $0 < \lambda < 1/C_2$, *where the expectation* $\mathbb{E}$ *is with respect to* y_{it} *given* z_{it}^0, $\|x_{it}\| \leq C_1$, *and* $\inf_{i,T} \sum_{t=1}^T (x_{it}^{(k)})^2/T \geq C_3$.

(v). (Non-collinearity) There exists $c > 0$ *such that for any* $\mathbf{v} = (v_1, \ldots, v_K) \in \mathbb{R}^{N \times K}$ *and* $\boldsymbol{\xi} \in \mathbb{R}^T$,

$$\frac{1}{NT} \operatorname{Tr}\left(\mathcal{M}_{(\boldsymbol{\alpha}^0, \mathbf{1}_N),} (\mathbf{v} \cdot \mathbf{X}) \mathcal{M}_{(\mathbf{1}_T, \boldsymbol{\xi})} (\mathbf{v} \cdot \mathbf{X})'\right) \geq c \|\mathbf{v}\|_{\max}^2$$

with probability one, where $\mathcal{M}_A = \mathbf{I} - A(A'A)^{-1}A'$ *is the coprojection matrix corresponding to column vectors in* A, $\mathbf{v} \cdot \mathbf{X} = \sum_{k=1}^K Diag(v_k) X_k$ *with* $X_k = (x_{it}^{(k)})_{i=1,\ldots,N; t=1,\ldots,T}$, *and* $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$.

(vi). (Concavity) For all N, T, *the function* $z \mapsto \ell_{it}(z)$ *is strictly concave over* $\mathcal{Z}$ *a.s. Furthermore, there exist positive constants* $b_{\min}$ *and* $b_{\max}$ *such that for all* $z \in \mathcal{Z}$, $b_{\min} \leq -\partial_{z^2} \ell_{it}(z) \leq b_{\max}$ *a.s. uniformly over* i, t, N, T.

Assumption 2.7.1 ressembles Assumption 1 in Chen et al. (2021). The main difference lies on Assumptions 2.7.1(iv) and 2.7.1(v). Assumption 2.7.1(iv) requires the score to have thin tails (sub-exponential) and is satisfied in many routinely used models, e.g., y_{it} has bounded support (binary, multimodal outcome) or y_{it} follows Poisson distribution. Assumption 2.7.1(v) adapts the non-collinearity condition to the setting with individual-specific slopes. Along the lines of their non-collinearity condition (as well as those in the existing literature, such as Bai (2009), Moon and Weidner (2015, 2017)), it requires the covariates to have sufficient variation once the fixed effects are partialled out and rules out covariates that do not vary across time or across individuals. Differently, due to the dimensionality in the number of slope parameters that increases asymptotically (proportionally to N), we instead use $\|\cdot\|_{\max}$ rather than $\|\cdot\|_{\mathrm{F}}$.[42]

The following proposition summarizes our consistency result in the presence of heterogeneous β_i. We adapt the proof of Lemma 1 in Chen et al. (2021) and the details can be found in Online Appendix 2.7.10.

Proposition 2.7.3 *Let Assumption 2.7.1 hold. Then, as* $N, T \to \infty$, *we have*

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_{\max} = O_P(N^{-3/8}).$$

Proposition 2.7.3 implies that the plug-in estimators of moments of β_i are consistent. To see this, suppose that β_i is a scalar i.i.d. random variable and the estimator

[42] The dimension of $\mathbf{v}$ is $N \times K$ and $\|\mathbf{v}\|_{\mathrm{F}}^2$ could be of order $O(N)$, while it is $O(1)$ in Chen et al. (2021) with $\mathbf{v} \in \mathbb{R}^{1 \times K}$. As a result, adopting $\|\mathbf{v}\|_{\mathrm{F}}^2$ in Assumption 2.7.1(v) may lead to a violation that does not occur in their framework. Instead, $\|\mathbf{v}\|_{\max}^2$ is still of order $O(1)$ and therefore immune to such violations.

of its k^{th} moments is $\widehat{m}_\beta = \frac{1}{N}\sum_{i=1}^N \widehat{\beta}_i^k$. Moreover, suppose that the compact set corresponding to β_i in (2.7.19), $\mathbf{B}_i$, is a subset of $[-\ln N, \ln N]$ for $i \leq N$, and $\mathbb{P}(\beta_i^0 \in \mathbf{B}_i, i = 1, \ldots, N|\mathbf{X}_N) \to 1$.[43] Then,

$$\begin{aligned}\left|\widehat{m}_\beta - \mathbb{E}\left[\beta_i^k\right]\right| &\leq \left|\widehat{m}_\beta - \frac{1}{N}\sum_{i=1}^N \beta_i^k\right| + \left|\frac{1}{N}\sum_{i=1}^N \beta_i^k - \mathbb{E}\left[\beta_i^k\right]\right| \\ &= \frac{1}{N}\sum_{i=1}^N \left(\left|\widehat{\beta}_i - \beta_i\right| \times \sum_{h=0}^{k-1}\left|\widehat{\beta}_i^h \beta_i^{k-h-1}\right|\right) + \left|\frac{1}{N}\sum_{i=1}^N \beta_i^k - \mathbb{E}\left[\beta_i^k\right]\right|.\end{aligned}$$

and, with probability approaching one,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \left(\left|\widehat{\beta}_i - \beta_i\right| \times \sum_{h=0}^{k-1}\left|\widehat{\beta}_i^h \beta_i^{k-h-1}\right|\right) &\leq k(\ln N)^{k-1}\frac{1}{N}\sum_{i=1}^N \left|\widehat{\beta}_i - \beta_i\right| \\ &\leq k(\ln N)^{k-1}\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_{\max} \\ &\to 0.\end{aligned}$$

Moreover, according to the law of large numbers, $\left|\frac{1}{N}\sum_{i=1}^N \beta_i^k - \mathbb{E}\left[\beta_i^k\right]\right| \xrightarrow{p} 0$. Then, we obtain that $\left|\widehat{m}_\beta - \mathbb{E}\left[\beta_i^k\right]\right| \xrightarrow{p} 0$. Furthermore, it is straightforward to show that $\widehat{m}_\beta - \mathbb{E}\left[\beta_i^k\right] = O_P(N^{-\frac{3}{8}}(\ln N)^{k-1})$ (and $O_P(N^{-\frac{3}{8}})$ when β_i are bounded).

2.7.4 Monte Carlo Experiments: Details

In our Monte Carlo simulations, the results are obtained by using 50 replications. The **RMSE to MLE** is the average Root Mean Squared Error to the joint maximum likelihood estimator (MLE) is defined as:

$$\text{RMSE}(\widehat{\theta}, \widehat{\theta}^{\text{MLE}}) := \frac{1}{50}\sum_{b=1}^{50}\sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(\widehat{\theta}_j^{(b)} - \widehat{\theta}_j^{\text{MLE}(b)}\right)^2}, \tag{2.7.20}$$

where 50 is the number of Monte Carlo repetitions, $\widehat{\theta}$ is a $d-$dimensional estimator. The APEs $\widehat{\delta}_{NT}$ is defined as:

$$\widehat{\delta}_{NT} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widehat{\beta}_{NT}\Lambda'(x_{it}\widehat{\beta}_{NT} + \widehat{\alpha}_{NT,i} + \widehat{\xi}_{NT,t}). \tag{2.7.21}$$

FPMLE and FPMLE^{++} are implemented by our Python package `nlmfe`. For both algorithms and the MLE, we initialize $(\alpha^{(0)\prime}, \xi^{(0)\prime}, \beta^{(0)\prime})' = \mathbf{0}_{N+T}$ (results are not sensitive to this choice). FPMLE^{++} employs either a constant step size of $\nu^{(s)} \approx 1/(NT)$ or an Hessian step. FPMLE employs the Newton Conjugate Gradient method implemented in the `minimize()` function from the Python class `scipy.optimize`. Moreover, in both the Monte Carlo experiments and empirical applications, besides the

[43] This condition accommodates the cases of bounded and unbounded (with thin tail such as Gaussian) β_i.

number of iterations, we jointly use a stopping criterion based on the variation of the objective function generated by the previous iterate (e.g., the iteration stops as soon as this variation is less than 10^{-5}). For the MLE, we compute it using the `LogisticRegression`(`penalty`='none', `tol`=1e-4, `C`=1.0, `fit_intercept`=False, `solver`='newton-cg', `max_iter`=1000) function from the `sklearn.linear_model` Python class. The **CPU time** is computed by Python's `time.perf_counter()` and measures the average user CPU time (in seconds) for the estimation with a Microsoft Windows 10 Enterprise laptop Intel(R) Core(TM) i7-1165G7MQ CPU @ 2.80GHz 1.69 GHz, 16GB RAM.

TABLE 2.2: Numerical Convergence – Logit Model with Homogeneous Slope ($N = T = 200$)

		FPMLE^{++}		**FPMLE**	
		RMSE to MLE		**RMSE to MLE**	
DGP	**#iter**	$\widehat{\beta}^{++}$	$\widehat{\delta}^{++}$	$\widehat{\beta}$	$\widehat{\delta}$
i.	1	0.18381	0.02711	0.01918	0.00178
	3	0.00069	0.00005	0.00007	0.00001
	20	0.00005	0.00001	0.00006	0.00001
ii.	1	0.13534	0.01783	0.01999	0.00187
	3	0.00043	0.00003	0.00008	0.00001
	20	0.00005	0.00001	0.00007	0.00001
iii.	1	0.35989	0.06817	0.19461	0.04555
	3	0.02415	0.00504	0.01096	0.00238
	20	0.00009	0.00002	0.00024	0.00005
iv.	1	0.30205	0.05693	0.12703	0.03163
	3	0.00943	0.00200	0.00257	0.00058
	20	0.00006	0.00001	0.00015	0.00003

Notes: For each DGP, each row reports the results obtained after **#iter** iterations and based on 50 replications.

Poisson Count Model with Heterogeneous Slopes

Consider a static Poisson count model with heterogeneous slopes: for $y = 0, 1, \ldots,$

$$\mathbb{P}(y_{it} = y|(x_{is})_{s=1}^{t}, \alpha_i, \xi_t, \beta_i^0) = \frac{\exp(y(x_{it}\beta_i^0 + \alpha_i + \xi_t))\exp(-\exp(x_{it}\beta_{0,i} + \alpha_i + \xi_t))}{y!},$$

with $\alpha_1 = 0$, $(\alpha_i)_{2\leq i\leq N} \overset{iid}{\sim} \mathcal{N}(0, 1/16), (\xi_t)_{1\leq t\leq T} \overset{iid}{\sim} \mathcal{N}(0, 1/16)$, and $\beta_{0,i} \overset{iid}{\sim} \mathcal{N}(1, 1/10)$.

2.7.5 Empirical Illustrations: Additional Results

TABLE 2.3: Inference – Poisson Model with Heterogeneous Slopes

	$\widehat{\mathbb{E}}(\beta_{0i})$			$\sqrt{\widehat{\mathbb{V}}(\beta_{0i})}$			$\widehat{\text{APE}}$		
Quantile interval (%)	[2.5, 97.5]	[4, 99]	[1, 96]	[2.5, 97.5]	[4, 99]	[1, 96]	[2.5, 97.5]	[4, 99]	[1, 96]
DGP									
i.	.9600	.9350	.9820	.6910	.7920	.6220	.9780	.9780	.9760
ii	.9820	.9660	.9900	.6370	.7570	.5370	.9920	.9860	.9970
iii.	.8560	.8270	.8830	.8310	.8590	.8030	.5530	.5940	.7660
iv.	.8850	.8510	.9190	.6430	.8050	.6740	.5970	.6390	.7900

Notes: Data are generated from the Poisson model described in Appendix 2.7.4 with $N = T = 50$. The coverages are computed based on 1,000 replications. For each repetition, we implement percentile Bootstrap jackknife CI's based on 200 Bootstrap samples. All computations are performed with FPMLE^{++} with at most 2 Hessian step iterations.

TABLE 2.4: Regressions of $-\widehat{\gamma}_j^{\text{exp}}$ and $-\widehat{\gamma}_i^{\text{imp}}$ over Observed Characteristics of a Country

	WTO member	Island country	Landlocked	Constant	R^2
$-\widehat{\gamma}_j^{\text{exp}}$	0.067	−0.024	−0.023	−0.040	20.93%
	(0.011)	(0.014)	(0.016)	(0.010)	
$-\widehat{\gamma}_i^{\text{imp}}$	0.034	−0.004	−0.004	−0.033	13.01%
	(0.007)	(0.009)	(0.010)	(0.006)	

Notes: Both regressions are implemented based on 157 estimated $-\widehat{\gamma}_i^{\text{exp}}$ and 158 estimated $-\widehat{\gamma}_j^{\text{imp}}$.

TABLE 2.5: Regressions of $\widehat{\beta}_i$ and $\widehat{\eta}_i$ over Observed Characteristics of Firm i

	Sales	R&D Exp.	Tobin's Q	Sector dummies	R^2
$\hat{\beta}_i$	0.0011	0.0003	0.0005	Yes	38.27%
	(0.0004)	(0.0003)	(0.0002)		
$\hat{\eta}_i$	0.0001	0.00005	−0.00004	Yes	11.02%
	(0.0001)	(0.00006)	(0.00004)		

Notes: Both regressions are implemented based on 452 estimated $\hat{\beta}_i$ and $\hat{\eta}_i$. Regressors are defined as the average of their values across the time period in the data. Sector dummies are defined using variable *sic4*.

TABLE 2.6: Correlations and Variance Decomposition

	$(\hat{\eta}_i, \hat{\beta}_i)$	$(z_i\hat{\gamma}^{\eta}, z_i\hat{\gamma}^{\beta})$	$(\hat{\zeta}_i^{\eta}, \hat{\zeta}_i^{\beta})$
Corr.	23.08%	54.36%	16.09%
Co-Variance Decom.	100%	48.84%	51.16%

Notes: The variance decomposition is based on (2.5.5) and (2.5.6). The observed characteristics z_i are the same as in Table 2.5.

2.7.6 Existence and Uniqueness of Coordinate-Wise Minima (Proof of Theorem 2.3.2)

Without loss of generality, let us focus on minimization of $f_{N+T,\alpha,\xi}$ over $\mathcal{B}$ and drop the indices (α, ξ) for convenience.

Existence. Let us denote $m = \inf_{b\in\mathcal{B}} f_{N+T}(b)$ and define

$$\mathcal{B}_0 = \begin{cases} \{b \in \mathcal{B} : f_{N+T}(b) \leq m+1\} & \text{if } m \in \mathbb{R}, \\ \{b \in \mathcal{B} : f_{N+T}(b) \leq 0\} & \text{if } m = -\infty. \end{cases}$$

By Assumption 2.3.2(ii), $\lim_{\|\theta\|\to\infty} \mathscr{L}_{NT}(\theta) = -\infty$ which implies $\lim_{\|b\|\to\infty} f_{N+T}(b) = +\infty$, i.e., f_{N+T} is coercive. Hence, $\mathcal{B}_0$ is bounded (if not, we would have $(b_n)_n \subset \mathcal{B}_0$ such that $\|b_n\| \to +\infty$ and thus $f_{N+T}(b_n) \to \infty$ whereas, for all n, $f_{N+T}(b_n) \leq \max(m+1, 0)$). Next, since f_{N+T} is continuous, $\mathcal{B}_0$ is a closed subset of $\mathcal{B}$ and, because $\mathcal{B}$ is closed, $\mathcal{B}_0$ is itself a closed subset of $\mathbb{R}^K$. As a closed and bounded set of $\mathbb{R}^K$, $\mathcal{B}_0$ is compact in $\mathbb{R}^K$. By Weirstrass theorem, f_{N+T} reaches $\inf_{b\in\mathcal{B}_0} f_{N+T}(b)$. Let $b^* \in \mathcal{B}^0$ such that $f_{N+T}(b^*) = \inf_{b\in\mathcal{B}_0} f_{N+T}(b)$. Let us show that b^* is a minimum of f_{N+T} over $\mathcal{B}$. Let $b \in \mathcal{B}$. If $b \in \mathcal{B}_0$, $f_{N+T}(b^*) \leq f(b)$ by definition of b^*. If $b \notin \mathcal{B}_0$ then, either $m \in \mathbb{R}$ and thus $f_{N+T}(b) \geq m+1 \geq f_{N+T}(b^*)$, either $m = -\infty$ and thus $f_{N+T}(b^*) \leq 0 \leq f_{N+T}(b)$. In both cases, $f_{N+T}(b) \geq f_{N+T}(b^*)$.

Uniqueness. Assume that $B^* := \arg\min_{b\in\mathcal{B}} f_{N+T}(b)$ has more than one point. Let b_1 and b_2 be two distinct solutions, *i.e.*, $f_{N+T}(b_1) = f_{N+T}(b_2) = \overline{f}_{N+T}$ and $b_1 \neq b_2$. By Assumption 2.3.2(ii), $\theta \mapsto \mathscr{L}_{NT}(\theta)$ is strictly concave, and therefore f_{N+T} is strictly convex. Since f_{N+T} is convex, the set B^* is also convex. Hence, for any $t \in (0,1)$, $tb_1 + (1-t)b_2 \in B^*$ and thus

$$f_{N+T}(tb_1 + (1-t)b_2) = \overline{f}_{N+T}. \tag{2.7.22}$$

By strict convexity of f_{N+T} over $\mathcal{B}$ and since $B^* \subset \mathcal{B}$, we have

$$f_{N+T}(tb_1 + (1-t)b_2) < tf_{N+T}(b_1) + (1-t)f_{N+T}(b_2) = \overline{f}_{N+T},$$

which contradicts Equation (2.7.22). Hence f_{N+T} has a unique minimum over $\mathcal{B}$. We note that the same reasoning applied to f shows the existence and uniqueness of $\widehat{\theta}^{\text{MLE}}_{NT}$.

2.7.7 Extension of Theorem 2.3.1 to Multimodal Outcomes

Consider a model with multimodal outcome: the probability of individual i from choosing $y_{it} \in \{1, \ldots, J\}$ at time t is

$$\mathbb{P}(y_{it} = j | (\alpha_{ij}, \xi_{sj}, \beta_{ij}, x_{isj})_{j=1,\ldots,J}^{s=1,\ldots,t}) = g_j(v_{it}), \tag{2.7.23}$$

where $v_{it} = (v_{itj})_{j=1}^J$ with $v_{itj} = \alpha_{ij} + \xi_{tj} + x'_{itj}\beta_{ij}$, $\sum_{j=1}^J g_j(v_{it}) < 1$, and J is known. The residual probability, $g_0(v_{it}) = 1 - \sum_{j=1}^J g_j(v_{it})$, is usually defined as

the probability of choosing the outside option. Model (2.7.23) is a common setting in empirical research such as demand estimation (Berry et al., 2013; Dubois et al., 2020). Define $g(v_{it}) = (g_j(v_{it}))_{j=1}^J$, $\alpha_i = (\alpha_{ij})_{j=1}^J$, $\beta_i = (\beta_{ij})_{j=1}^J$, $\xi_t = (\xi_{tj})_{j=1}^J$. For $j = 1, \ldots, J$, we normalize $\alpha_{1j} = 0$, $\xi_{1j} = 0$, and $\beta_{1j}^{(1)} = 1$. We aim to identify $(\alpha_i, \beta_i)_{i\in\mathbf{N}}$, $(\xi_t)_{t\in\mathbf{T}}$, and $g(\cdot)$. Similarly to the single-index case, we define a compensating vector of dimension J:

$$z_i(x^{(1)}; x^{(2)}) = (z_{ij}(x^{(1)}; x^{(2)}))_{j=1}^J = \left(\alpha_{ij} + x_{ij}^{(1)}\beta_{ij}^{(1)} + x_{ij}^{(2)}(\beta_{ij}^{(2)} - \beta_{1j}^{(2)})\right)_{j=1}^J \quad (2.7.24)$$

$z_i(x^{(1)}; x^{(2)})$ is the needed value of $x^{(1)}$ for individual 1 with $x^{(2)}$ to make her and i's indices equal: for $j = 1, \ldots, J$,

$$\alpha_{1j} + \xi_{tj} + \beta_{1j}^{(1)} z_{ij}(x^{(1)}; x^{(2)}) + \beta_{1j}^{(2)} x_{1j}^{(2)} = \alpha_{ij} + \xi_{tj} + \beta_{ij}^{(1)} x_{ij}^{(1)} + \beta_{ij}^{(2)} x_{ij}^{(2)}.$$

The definition of compensation in Definition 2.3.1 can accordingly be extended. For $v = (v_1, \ldots, v_J) \in \mathbb{R}^{2\times J}$, let $\mathrm{P}^J_{(1,\beta_1^{(2)})}(v) := \left(\mathrm{P}_{(1,\beta_{11}^{(2)})}(v_1), \ldots, \mathrm{P}_{(1,\beta_{1J}^{(2)})}(v_J)\right)$. We now extend Assumption 2.3.1 to the following:

Assumption 2.7.2

(i). The mapping g satisfies the following conditions:

- *(a) The support of $(v_{it1}, \ldots, v_{itJ})$, $\mathcal{V}$, is a Cartesian product.*
- *(b) (Weak substitutes) $g_j(v)$ is weakly decreasing in v_k for all $j = 1, \ldots, J$ and $k \notin \{0, j\}$.*
- *(c) (Connected strict substitution) For all $v \in \mathcal{V}$ and any nonempty subset of $\{1, \ldots, J\}$, $\mathcal{K}$, there exists $k \in \mathcal{K}$ and $l \notin \mathcal{K}$ such that g_l is strictly decreasing in v_k.*

(ii). (a) For all $i \in \mathbf{N}$, conditional on (α_i, β_i), $\{(y_{it}, x_{it1}, \ldots, x_{itJ})\}_{t\geq 2}$ is a strictly stationary and strong mixing process with mixing coefficients τ_t that satisfy $\tau_t \leq C\rho^t$.

- *(b) For $t \in \mathbf{T}$, conditional on ξ_t, $\{(y_{it}, x_{it1}, \ldots, x_{itJ}, \alpha_i, \beta_i)\}_{i\in\mathbf{N}}$ are independent.*

(iii). For all $(i, i', t, j) \in \mathbf{N}^2 \times \mathbf{T} \times \{1, \ldots, J\}$, ξ_t is independent of $(\alpha_{ij}, \beta_{ij}, x_{itj}^{(1)})$ conditional on $x_{itj}^{(2)}$. Moreover, $\xi_t | \left\{x_{itj}^{(2)} = x^{(2)}\right\} \overset{d}{=} \xi_t | \left\{x_{i'tj}^{(2)} = x^{(2)}\right\} \sim F_\xi(\xi; x^{(2)})$.

(iv). Any individual $i \in \mathbf{N}$ is compensable by individual 1 at least at $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}_i$ for $k = 1, 2, 3$ with

$$\begin{bmatrix} 1 & x_j^{(1)1} & x_j^{(2)1} \\ 1 & x_j^{(1)2} & x_j^{(2)2} \\ 1 & x_j^{(1)3} & x_j^{(2)3} \end{bmatrix}$$

being nonsingular for $j = 1, \ldots, J$.

(v). Denote $\mathcal{Z}_t = \cap_{i\in\mathbf{N}}\mathcal{Z}_{it}$ *and* $\mathcal{Z}_{jt}$ *the projection of* $\mathcal{Z}_t$ *along its* j^{th} *coordinate.*

(a) For any $j = 1,\ldots,J$, *there exists some* $(t, r_j) \in \mathbf{T}\times\mathbb{R}$, *such that* $\{(z_j, x_j^{(2)}) \in \mathcal{Z}_{jt} : \mathrm{P}_{(1,\beta_{1j}^{(2)})}(z_j, x_j^{(2)}) = r_j\}$ *is not a singleton.*

(b) For all $t \in \mathbf{T}$, $\left(\cap_{i\in\mathbf{N}}\mathrm{P}^J_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it}) + \xi_t\right) \cap \left(\cap_{i\in\mathbf{N}}\mathrm{P}^J_{(1,\beta_1^{(2)})}(\mathcal{Z}_{i1})\right) \neq \emptyset$.

Assumption 2.7.2(i) is a multi-index version of Assumption 2.3.1(i). It is motivated by sufficient conditions for the invertibility of demand by Berry et al. (2013) and usually satisfied in the setting of discrete-choice random utility models with separably additive index and idiosyncratic error in indirect utility. This assumption implies that g is a bijection from $\mathcal{V}$ to $g(\mathcal{V})$. Moreover, it implies that the aggregated choice probability function, i.e., the integral of $g(v_{it})$ over ξ_t for a given i, satisfies Assumption 2.7.2(i) and is therefore a bijection. Both bijection properties enable to apply the argument of compensating variable as in the single-index case. As argued in Berry et al. (2013), Assumption 2.7.2(i) is convenient in practice due to its Cartesian support requirement and it applies even when g may not be not differentiable. This contrasts other arguments such as those by Gale and Nikaido (1965) which require rectangular support condition and differentiability of g. In contrast, due to the weak substitutes requirement in Assumption 2.7.2(i)b, the derivative of g_j (if differentiable) with respect to v_k is restricted to be nonpositive for all $k \neq j$. As an alternative, one can require $g_j(\cdot)$ to be strictly increasing with respect to index v_j for all $j = 1,\ldots,J$ and the mapping g to have strictly diagonally dominant Jacobian, which will also imply the bijection properties we need to apply the argument of compensating variable but allows for positive cross derivatives in g. Assumptions 2.7.2(ii)-(v) are similar to those in Assumption 2.3.1 and are accommodated to the fact that the compensating vector is of dimension J.

Theorem 2.7.4 *Suppose that Assumptions 2.7.2(i)-(iv) hold.*

- $\beta_i^{(1)}$, α_i, *and* $\beta_i^{(2)} - \beta_1^{(2)}$ *are identified for* $i \in \mathbf{N}$.
- *If Assumption 2.7.2(v) further holds, then*
  - ξ_t *and* $\beta_i^{(2)}$ *are identified for* $i \in \mathbf{N}$ *and* $t \in \mathbf{T}$.
  - $g(y; v)$ *is identified for* $(y, v) \in \mathcal{Y}\times\cup_{t\in\mathbf{T}}\left(\cap_{i\in\mathbf{N}}\mathrm{P}^J_{(1,\beta_1^{(2)})}(\mathcal{Z}_{it}) + \xi_t\right)$.

First, for $i \in \mathbf{N}$ and $x \in \mathcal{X}_i$, we identify the following vector of quantities using Assumptions 2.7.2(ii)a and (iii):

$$G_i(x) := (\Gamma_{ij}(x))_{j=1}^J,$$
$$\Gamma_{ij}(x) := \mathbb{E}\left[\mathbf{1}\{y_{it} = j\}|x_{itj} = x, \alpha_{ij}, \beta_{ij}\right] = \int g_j(\alpha_{ij} + x'\beta_{ij} + \xi)dF_\xi(\xi; x^{(2)})d\xi. \tag{2.7.25}$$

To apply the argument of compensating variable in the proof for the single-index case, we need first to prove that g and G_i are bijections from $\mathcal{V}$ and the support of

$\alpha_i + x'\beta_i$ (which is supposed to be a Cartesian product) to their images, respectively. The former bijectivity is immediately implied by Assumption 2.7.2(i) using the arguments in Berry et al. (2013). To prove the latter injectivity, it suffices to verify that G_i satisfies the three requirements in Assumption 2.7.2(i). The first and second requirements are immediate because of the definition of G_i. Moreover, because of the weak substitutes of G_i and connected strict substitution of g in the integral for any ξ, the third requirement holds as well.

Given the bijectivities of g and G_i, we can now apply the argument of compensating variable using Assumptions 2.7.2(ii)-(v) and z_i defined in (2.7.24). The proof is essentially the same as that of Theorem 2.3.1.

2.7.8 Heterogeneous Slope Across Time

We give sufficient conditions for the identification of model (2.2.2) with heterogeneous slopes across time periods. To start with, define a compensating variable:

$$z^t(x^{(1)}; x^{(2)}) = \xi_t + \beta_t^{(1)} x^{(1)} + x^{(2)}(\beta_t^{(2)} - \beta_1^{(2)}) \tag{2.7.26}$$

Intuitively, $z^t(x^{(1)}; x^{(2)})$ is the needed value of $x^{(1)}$ for individual i with $x^{(2)}$ at $t = 1$ to make her indices at time 1 and t equal: $\alpha_i + \xi_1 + \beta_1^{(1)} z^t(x^{(1)}; x^{(2)}) + \beta_1^{(2)} x^{(2)} = \alpha_i + \xi_t + \beta_t^{(1)} x^{(1)} + \beta_t^{(2)} x^{(2)}$.

Definition 2.7.2 *Time period t is compensable at $(x^{(1)}, x^{(2)}) \in \mathcal{X}^t := Supp(x_{it}|\xi_t, \beta_t)$ by time period 1 if $(z^t(x^{(1)}; x^{(2)}), x^{(2)}) \in \mathcal{X}^1$.*[44]

Let $\mathcal{Z}^t = \left\{(z^t(x^{(1)}; x^{(2)}), x^{(2)}) : (x^{(1)}, x^{(2)}) \in \mathcal{X}^t\right\}$. Denote by $\mathcal{Z}^{ti}$ the support of $(z^t(x_{it}^{(1)}; x_{it}^{(2)}), x_{it}^{(2)})$ conditional on α_i.

Assumption 2.7.3

(i). There exists $y \in \mathcal{Y}$ such that the function $g(y; v)$ is strictly monotonic in v.

(ii). (a) For all $t \in \mathbf{T}$, conditional on (β_t, ξ_t), $\{(y_{it}, x_{it})\}_{i \in \mathbf{N}}$ are independent.

(b) For all $i \in \mathbf{N}$, conditional on α_i, $\{(y_{it}, x_{it}, \beta_t, \xi_t)\}_{t \geq 2}$ is a strictly stationary and strong mixing process with mixing coefficients τ_t that satisfy $\tau_t \leq C\rho^t$.

(iii). For all $(i, i', t) \in \mathbf{N}^2 \times \mathbf{T}$, α_i is independent of $(\xi_t, \beta_t, x_{it}^{(1)})$ conditional on $x_{it}^{(2)}$. Moreover, $\alpha_i | \left\{x_{it}^{(2)} = x^{(2)}\right\} \overset{d}{=} \alpha_{i'} | \left\{x_{i't}^{(2)} = x^{(2)}\right\} \sim F_\alpha(\alpha; x^{(2)})$.

(iv). Any time period $t \in \mathbf{T}$ is compensable by time period 1 at least at $(x^{(1)k}, x^{(2)k}) \in \mathcal{X}^t$ for $k = 1, 2, 3$ with

$$\begin{bmatrix} 1 & x^{(1)1} & x^{(2)1} \\ 1 & x^{(1)2} & x^{(2)2} \\ 1 & x^{(1)3} & x^{(2)3} \end{bmatrix}$$

being nonsingular.

[44] We assume that $\mathcal{X}^t$ does not depend on i to simplify the exposition. This holds, e.g., if $\{x_{it}\}_{i \geq 2}$ is i.i.d. conditional on (ξ_t, β_t) for all $t \in \mathbf{T}$.

(v). Denote $\mathcal{Z}^i = \cap_{t\in\mathbf{T}}\mathcal{Z}^{ti}$.

(a) For some $(i,r) \in \mathbf{N}\times\mathbb{R}$, $\left\{(z,x^{(2)}) \in \mathcal{Z}^i : \mathrm{P}_{(1,\beta_1^{(2)})}(z,x^{(2)}) = r\right\}$ *is not a singleton.*

(b) For all $i \in \mathbf{N}$, $\left(\cap_{t\in\mathbf{T}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}^{ti}) + \alpha_i\right) \cap \left(\cap_{t\in\mathbf{T}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}^{t1})\right) \neq \emptyset$.

Theorem 2.7.5 *Suppose that Assumptions 2.7.3(i)-(iv) hold.*

- $\beta_t^{(1)}$, ξ_t, *and* $\beta_t^{(2)} - \beta_1^{(2)}$ *are identified for* $t \in \mathbf{T}$.
- *If Assumptions 2.7.3(v) further holds, then*
  - α_i *and* $\beta_t^{(2)}$ *are identified for* $i \in \mathbf{N}$ *and* $t \in \mathbf{T}$.
  - $g(y;v)$ *is identified for* $(y,v) \in \mathcal{Y}\times\cup_{i\in\mathbf{N}}\left(\cap_{t\in\mathbf{T}}\mathrm{P}_{(1,\beta_1^{(2)})}(\mathcal{Z}^{ti}) + \alpha_i\right)$.

Assumption 2.7.3, as well as $z^t(x^{(1)};x^{(2)})$, mirrors Assumption 2.3.1 and $z_i(x^{(1)};x^{(2)})$ regarding the individual and time dimensions. Consequently, one can alter the roles of the two dimensions in the proof of Theorem 2.3.1 to show Theorem 2.7.5. Apart from this difference, the proofs are essentially the same.

2.7.9 Extension of FPMLE and FPMLE^{++}

Heterogeneous β_i

We describe the extension to the case of heterogeneous slopes β_i. The extension to the case of β_t is similar.

Fully Heterogeneous β_i**.** In this case, all the components of β_i are individual-i specific. Let $N\times K$ matrix $\boldsymbol{\beta}^0 \in \mathcal{B}^N$ denote heterogeneous slopes $(\beta_1^0,\ldots,\beta_N^0)'$. We introduce an additional step in FPMLE and FPMLE^{++} to update each slope.

Algorithm FPMLE (Fully Heterogeneous Slopes):

1. Let $(\xi_1^{(0)},\ldots,\xi_T^{(0)},(\boldsymbol{\beta}^{(0)})')' \in \Xi\times\mathcal{B}^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1,2,\ldots\}$. Set $s = 0$.

2 Compute (in parallel) for all $i \in \{2,\ldots,N\}$:

$$\alpha_i^{(s+1)} \in \arg\max_{\alpha\in\mathcal{A}_i} \sum_{t=1}^{T} \log g(y_{it};\alpha + x_{it}'\beta_i^{(s)} + \xi_t^{(s)}),$$

and

$$\beta_i^{(s+1)} \in \arg\max_{\beta\in\mathcal{B}} \sum_{t=1}^{T} \log g(y_{it};\alpha_i^{(s+1)} + x_{it}'\beta + \xi_t^{(s)}).$$

3. Compute (in parallel) for all $t \in \{1,\ldots,T\}$:

$$\xi_t^{(s+1)} \in \arg\max_{\xi\in\Xi_t} \sum_{i=1}^{N} \log g(y_{it};\alpha_i^{(s+1)} + x_{it}'\beta_i^{(s+1)} + \xi).$$

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

To use this algorithm, one can set (`het_exog=range`(K), `fast=False`) in the `TwoWayFPMLE` class from our `nlmfe` package.

Algorithm FPMLE^{++} (Fully Heterogeneous Slopes):

1. Let $(\alpha_2^{(0)}, \ldots, \alpha_N^{(0)}, \xi_1^{(0)}, \ldots, \xi_T^{(0)}, (\boldsymbol{\beta}^{(0)})')' \in \mathcal{A} \times \Xi \times \mathcal{B}^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, \ldots\}$. Let $\left\{\nu^{(s)}\right\}_{s \geq 0}$ be some bounded sequence of positive scalars such that $\liminf_s \nu^{(s)} > 0$. Set $s = 0$.

2 Compute:

$$\begin{pmatrix} \alpha_2^{(s+1)} \\ \vdots \\ \alpha_N^{(s+1)} \end{pmatrix} = \left[\begin{pmatrix} \alpha_2^{(s)} \\ \vdots \\ \alpha_N^{(s)} \end{pmatrix} - \nu^{(s)} \begin{pmatrix} \sum_{t=1}^T \frac{g'}{g}(y_{2t}; \alpha_2^{(s)} + x_{2t}'\beta_2^{(s)} + \xi_t^{(s)}) \\ \vdots \\ \sum_{t=1}^T \frac{g'}{g}(y_{Nt}; \alpha_N^{(s)} + x_{Nt}'\beta_N^{(s)} + \xi_t^{(s)}) \end{pmatrix} \right]_{\mathcal{A}}^{+},$$

where $[v]_{\mathcal{A}}^{+}$ denotes the vector whose i-th coordinate is the orthogonal projection of v_i on $\mathcal{A}_i$.

3. Compute:

$$\begin{pmatrix} \xi_1^{(s+1)} \\ \vdots \\ \xi_T^{(s+1)} \end{pmatrix} = \left[\begin{pmatrix} \xi_1^{(s)} \\ \vdots \\ \xi_T^{(s)} \end{pmatrix} - \nu^{(s)} \begin{pmatrix} \sum_{i=1}^N \frac{g'}{g}(y_{i1}; \alpha_i^{(s+1)} + x_{i1}'\beta_i^{(s)} + \xi_1^{(s)}) \\ \vdots \\ \sum_{i=1}^N \frac{g'}{g}(y_{iT}; \alpha_i^{(s+1)} + x_{iT}'\beta_i^{(s)} + \xi_T^{(s)}) \end{pmatrix} \right]_{\Xi}^{+},$$

where $[v]_{\Xi}^{+}$ denotes the vector whose t-th coordinate is the orthogonal projection of v_t on Ξ_t.

4. Compute (keeping it vectorized) for all $i \in \{1, \ldots, N\}$:

$$\beta_i^{(s+1)} = \left[\beta_i^{(s)} - \nu^{(s)} \sum_{t=1}^T x_{it} \frac{g'}{g}(y_{it}; \alpha_i^{(s+1)} + x_{it}'\beta_i^{(s)} + \xi_t^{(s+1)}) \right]_{\mathcal{B}}^{+},$$

where $[v]_{\mathcal{B}}^{+}$ denotes the orthogonal projection of v on $\mathcal{B}$.

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

To use this algorithm, one can set (`het_exog=range`(K), `fast=True`) in the `TwoWayFPMLE` class from our `nlmfe` package.

Partly Heterogeneous β_i. In this case, some components in β_i are heterogeneous across individuals, while the other components in β_i are homogeneous. Let $H \subset \{1, \ldots, K\}$ be the subset indexing variables with heterogeneous coefficients and let $\boldsymbol{\beta}_H^0 \in \mathcal{B}_H^N$ with $\mathcal{B}_H \subset \mathbb{R}^{|H|}$ denote the true heterogeneous parameter values. Let $\beta_{H^c}^0 \in \mathcal{B}_{H^c}$ with $\mathcal{B}_{H^c} \subset \mathbb{R}^{K-|H|}$ denote the true homogeneous parameter values. For any subset $S \subset \{1, \ldots, K\}$ and vector $u \in \mathbb{R}^K$, let $u_S \in \mathbb{R}^{|S|}$ be the vector obtained

after "removing" coordinates $s \in S$ from u. We propose that following extensions of FPMLE and FPMLE^{++}.

Algorithm FPMLE (Partly Heterogeneous β_i):

1. Let $(\xi_1^{(0)}, \ldots, \xi_T^{(0)}, (\beta_{H^c}^{(0)})', (\boldsymbol{\beta}_H^{(0)})')' \in \Xi \times \mathcal{B}_{H^c} \times \mathcal{B}_H^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, \ldots\}$. Set $s = 0$.

2 Compute (in parallel) for all $i \in \{2, \ldots, N\}$:

$$\alpha_i^{(s+1)} \in \arg\max_{\alpha \in \mathcal{A}_i} \sum_{t=1}^{T} \log g(y_{it}; \alpha + x_{H^c,it}' \beta_{H^c}^{(s)} + x_{H,it}' \beta_{H,i}^{(s)} + \xi_t^{(s)}),$$

and

$$\beta_{H,i}^{(s+1)} \in \arg\max_{\beta \in \mathcal{B}_H} \sum_{t=1}^{T} \log g(y_{it}; \alpha_i^{(s+1)} + x_{H^c,it}' \beta_{H^c}^{(s)} + x_{H,it}' \beta + \xi_t^{(s)}).$$

3. Compute (in parallel) for all $t \in \{1, \ldots, T\}$:

$$\xi_t^{(s+1)} \in \arg\max_{\xi \in \Xi_t} \sum_{i=1}^{N} \log g(y_{it}; \alpha_i^{(s+1)} + x_{H^c,it}' \beta_{H^c}^{(s)} + x_{H,it}' \beta_{H,i}^{(s+1)} + \xi).$$

4. Compute:

$$\beta_{H^c}^{(s+1)} \in \arg\max_{\beta \in \mathcal{B}_{H^c}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log g(y_{it}; \alpha_i^{(s+1)} + x_{H^c,it}' \beta + x_{H,it}' \beta_{H,i}^{(s+1)} + \xi_t^{(s+1)}).$$

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

To use this algorithm, one can set (`het_exog`=$H-1$, `fast=False`) in the `TwoWayFPMLE` class from our `nlmfe` package.[45]

Algorithm FPMLE^{++} (Partly Heterogeneous β_i):

1. Let $(\alpha_2^{(0)}, \ldots, \alpha_N^{(0)}, \xi_1^{(0)}, \ldots, \xi_T^{(0)}, (\boldsymbol{\beta}^{(0)})')' \in \mathcal{A} \times \Xi \times \mathcal{B}^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, \ldots\}$. Let $\left\{\nu^{(s)}\right\}_{s \geq 0}$ be some bounded sequence of positive scalars such that $\liminf_s \nu^{(s)} > 0$. Set $s = 0$.

2 Compute:

$$\begin{pmatrix} \alpha_2^{(s+1)} \\ \vdots \\ \alpha_N^{(s+1)} \end{pmatrix} = \left[\begin{pmatrix} \alpha_2^{(s)} \\ \vdots \\ \alpha_N^{(s)} \end{pmatrix} - \nu^{(s)} \begin{pmatrix} \sum_{t=1}^{T} \frac{g'}{g}(y_{2t}; \alpha_2^{(s)} + x_{H^c,2t}' \beta_{H^c}^{(s)} + x_{H,2t}' \beta_{H,2}^{(s)} + \xi_t^{(s)}) \\ \vdots \\ \sum_{t=1}^{T} \frac{g'}{g}(y_{Nt}; \alpha_N^{(s)} + x_{H^c,Nt}' \beta_{H^c}^{(s)} + x_{H,Nt}' \beta_{H,N}^{(s)} + \xi_t^{(s)}) \end{pmatrix} \right]_{\mathcal{A}}^{+},$$

where $[v]_{\mathcal{A}}^{+}$ denotes the vector whose i-th coordinate is the orthogonal projection of v_i on $\mathcal{A}_i$, and

[45] The "-1" comes from Python's indexing system starting at 0.

3. Compute (keeping it vectorized) for all $i \in \{1, \dots, N\}$:

$$\beta_{H,i}^{(s+1)} = \left[\beta_{H,i}^{(s)} - \nu^{(s)} \sum_{t=1}^{T} x_{H,it} \frac{g'}{g}(y_{it}; \alpha_i^{(s+1)} + x'_{H^c,it}\beta_{H^c}^{(s)} + x'_{H,it}\beta_{H,i}^{(s)} + \xi_t^{(s)}) \right]_{\mathcal{B}_H}^{+},$$

where $[v]_{\mathcal{B}}^{+}$ denotes the orthogonal projection of v on $\mathcal{B}$.

3. Compute:

$$\begin{pmatrix} \xi_1^{(s+1)} \\ \vdots \\ \xi_T^{(s+1)} \end{pmatrix} = \left[\begin{pmatrix} \xi_1^{(s)} \\ \vdots \\ \xi_T^{(s)} \end{pmatrix} - \nu^{(s)} \begin{pmatrix} \sum_{i=1}^{N} \frac{g'}{g}(y_{i1}; \alpha_i^{(s+1)} + x'_{H^c,i1}\beta_{H^c}^{(s)} + x'_{H,i1}\beta_{H,i}^{(s+1)} + \xi_1^{(s)}) \\ \vdots \\ \sum_{i=1}^{N} \frac{g'}{g}(y_{iT}; \alpha_i^{(s+1)} + x'_{H^c,iT}\beta_{H^c}^{(s)} + x'_{H,iT}\beta_{H,i}^{(s+1)} + \xi_T^{(s)}) \end{pmatrix} \right]_{\Xi}^{+},$$

where $[v]_{\Xi}^{+}$ denotes the vector whose t-th coordinate is the orthogonal projection of v_t on Ξ_t.

4. Compute:

$$\beta_{H^c}^{(s+1)} = \left[\beta_{H^c}^{(s)} - \nu^{(s)} \sum_{t=1}^{T} x_{H^c,it} \frac{g'}{g}(y_{it}; \alpha_i^{(s+1)} + x'_{H^c,it}\beta_{H^c}^{(s)} + x'_{H,it}\beta_{H,i}^{(s+1)} + \xi_t^{(s+1)}) \right]_{\mathcal{B}_{H^c}}^{+}.$$

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

To use this algorithm, one can set (`het_exog`$=H-1$, `fast=True`) in the `TwoWayFPMLE` class from our `nlmfe` package.

Numerical Convergence without Concavity

In this appendix, we show that if FPMLE/FPMLE^{++} converges numerically, then it converges to a stationary point of the likelihood function (2.3.5). We prove it for the general case with a L-dimensional θ^g in g being estimated.

This property holds straightforwardly for FPMLE as long as the likelihood function is continuously differentiable. For FPMLE^{++}, we prove it in the next proposition.

Proposition 2.7.6 *Suppose that $\nu^{(s)} = \nu$, $\widehat{\theta}^{++(s)} \to \theta^*$, $\Theta_{NT} = \prod_{i=1}^{n_{\text{bloc}}} X_i$ is a product of convex compact sets X_i with 0 in their interior, and $\mathscr{L}_{NT}$ is continuously differentiable over $\mathbb{R}^{N+T+K+L-2}$. Then, $\frac{\partial \mathscr{L}_{NT}(\theta^*)}{\partial \theta} = 0$.*

Proof: By continuity of the orthogonal projection onto closed convex sets and continuous differentiability of $\mathscr{L}_{NT}(\cdot)$, we have

$$\theta_i^* = \left[\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*) \right]_{X_i}^{+}, \quad i = 1, \dots, n_{\text{blocs}}. \tag{2.7.27}$$

Fix $i \in \{1, \dots, n_{\text{blocs}}\}$ and let $L_i = \sup_{\theta \in \Theta_{NT}} \left\| \frac{\partial \mathscr{L}_{NT}(\theta)}{\partial \theta_i} \right\|$. Since $\mathscr{L}_{NT}$ is continuously differentiable and Θ_{NT} is compact, $0 \leq L_i < +\infty$. As $\theta_i^* \in X_i$ and X_i is bounded,

there exists $M_i > 0$ such that $\|\theta_i^*\| \leq M_i$. The triangle inequality yields

$$\left\|\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right\| \leq M_i + \nu L_i =: a_i,$$

where $a_i > 0$. Since 0 lies in the interior of X_i, there exists $b_i > 0$ sufficiently large such that $(1/b_i) \times \left(\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right) \in X_i$. By (2.7.27) and linearity of the orthogonal projection, it follows that

$$\begin{aligned}(1/b_i) \times \theta_i^* &= (1/b_i) \times \left[\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right]_{X_i}^+ \\ &= \left[(1/b_i) \times \left(\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right)\right]_{X_i}^+ \\ &= (1/b_i)\theta_i^* - (1/b_i)\frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*). \end{aligned} \tag{2.7.28}$$

The result then follows from (2.7.28) and $(1/b_i) \neq 0$ for all $i = 1, \ldots, n_{\text{bloc}}$. □

2.7.10 Proofs

Proof of Lemma 2.7.1

1. Let $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$. f is (μ, X)-strongly convex if and only if $g(y) \geq g(x) + \langle \nabla g(x), y - x\rangle$ for all $x, y \in X$, if and only if g is convex. By the monotone gradient condition for convexity, g is convex if and only if $\langle \nabla g(x) - \nabla g(y), x - y\rangle \geq 0$ for all $x, y \in X$, if and only if (2.7.5) holds.
2.(a) By twice differentiability of g, g is convex if and only if $\nabla^2 g \gtrsim 0$, if and only if $\nabla^2 f \gtrsim \mu I$. 2.(b) comes from an application of the fundamental theorem of calculus and the triangle inequality.

Proof of Proposition 2.7.3

For all $z_1, z_2 \in \mathcal{Z}$, the second-order Taylor expansion of $\ell_{it}(z_2)$ around z_1 yields

$$\ell_{it}(z_1) - \ell_{it}(z_2) = [\partial_z \ell_{it}(z_1)](z_1 - z_2) - \frac{1}{2}[\partial_{z^2} \ell_{it}(z^*)](z_1 - z_2)^2, \tag{2.7.29}$$

for some $z^* \in [\min\{z_1, z_2\}, \max\{z_1, z_2\}]$. Letting $e_{it} = \partial_z \ell_{it}/b_{\min}$ and $e := (e_{it})_{i,t}$, we have by definition of $\widehat{\boldsymbol{\theta}}$:

$$\begin{aligned} 0 \geq \mathcal{L}(\boldsymbol{\theta}^0) - \mathcal{L}(\widehat{\boldsymbol{\theta}}) &= \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T [\ell_{it}(z_{it}^0) - \ell_{it}(\widehat{z}_{it})] \\ &\geq \frac{b_{\min}}{2NT}\sum_{i=1}^N \sum_{t=1}^T (x_{it}'(\widehat{\beta}_i - \beta_i^0) + \widehat{\alpha}_i - \alpha_i^0 + \widehat{\xi}_t - \xi_t^0 - e_{it})^2 - \frac{b_{\min}}{2NT}\|e\|_{\text{F}}^2. \end{aligned}$$

Letting $\widehat{\boldsymbol{\lambda}} := (\widehat{\boldsymbol{\alpha}}, \mathbf{1}_N) \in \mathbb{R}^{N\times 2}$, $\widehat{\boldsymbol{f}} := (\mathbf{1}_T, \widehat{\boldsymbol{\xi}}) \in \mathbb{R}^{T\times 2}$, $\boldsymbol{\lambda}^0 := (\boldsymbol{\alpha}^0, \mathbf{1}_N)$, and $\boldsymbol{f}^0 := (\mathbf{1}_T, \boldsymbol{\xi}^0)$, we have

$$\begin{aligned}\frac{b_{\min}}{2NT}\|e\|_{\mathrm{F}}^2 &\geq \frac{b_{\min}}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}'(\widehat{\beta}_i-\beta_i^0)+\widehat{\alpha}_i-\alpha_i^0+\widehat{\xi}_t-\xi_t^0-e_{it})^2\\ &= \frac{b_{\min}}{2NT}\left\|\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}+\widehat{\boldsymbol{\lambda}}\widehat{\boldsymbol{f}}'-\boldsymbol{\lambda}^0\boldsymbol{f}^{0\prime}-e\right\|_{\mathrm{F}}^2.\end{aligned}$$

Because for any matrix A, $\|A\|_{\mathrm{F}}^2 = \mathrm{Tr}\,(AA')$, we then have

$$\begin{aligned}&\frac{1}{NT}\mathrm{Tr}\,(ee')\\ &\quad\geq \frac{1}{NT}\mathrm{Tr}\left[\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}+\widehat{\boldsymbol{\lambda}}\widehat{\boldsymbol{f}}'-\boldsymbol{\lambda}^0\boldsymbol{f}^{0\prime}-e\right)\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}+\widehat{\boldsymbol{\lambda}}\widehat{\boldsymbol{f}}'-\boldsymbol{\lambda}^0\boldsymbol{f}^{0\prime}-e\right)'\right].\end{aligned}$$

By using the same reasoning as Chen et al. (2021) p.313, we obtain

$$\tfrac{1}{NT}\mathrm{Tr}\,(ee') \geq \tfrac{1}{NT}\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}-e\right)\mathcal{M}_{\widehat{\boldsymbol{f}}}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}-e\right)'\right). \tag{2.7.30}$$

Let $\mathcal{P}_A = A(A'A)^{-1}A'$. First, note that:

$$\begin{aligned}\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}e\mathcal{M}_{\widehat{\boldsymbol{f}}}e'\right) &= \mathrm{Tr}\left(\left(\mathbf{I}-\mathcal{P}_{\boldsymbol{\lambda}^0}\right)e\left(\mathbf{I}-\mathcal{P}_{\widehat{\boldsymbol{f}}}\right)e'\right)\\ &= \mathrm{Tr}\,(ee') - \mathrm{Tr}\left(e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right) - \mathrm{Tr}\,(e'\mathcal{P}_{\boldsymbol{\lambda}^0}e)\\ &\quad + \mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right),\end{aligned}$$

and, using $\mathrm{Tr}(A) \leq \mathrm{rank}(A)\,\|A\|_2$,

$$\begin{aligned}\left|\mathrm{Tr}\left(e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right| &\leq \mathrm{rank}\left(e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\left\|e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right\|_2 \leq 2\|e\|_2^2,\\ \left|\mathrm{Tr}\,(e'\mathcal{P}_{\boldsymbol{\lambda}^0}e)\right| &\leq \mathrm{rank}\,(e'\mathcal{P}_{\boldsymbol{\lambda}^0}e)\left\|e'\mathcal{P}_{\boldsymbol{\lambda}^0}e\right\|_2 \leq 2\|e\|_2^2,\\ \left|\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right| &= \left|\mathrm{Tr}\left(\left[\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}\right]\left[\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}\right]'\right)\right| \leq 2\|e\|_2^2.\end{aligned} \tag{2.7.31}$$

Then, we obtain:

$$\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}e\mathcal{M}_{\widehat{\boldsymbol{f}}}e'\right) \geq \mathrm{Tr}\,(ee') - 6\|e\|_2^2. \tag{2.7.32}$$

Second, note that

$$\begin{aligned}&\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{M}_{\widehat{\boldsymbol{f}}}e'\right)\\ &= \mathrm{Tr}\left(\left(\mathbf{I}-\mathcal{P}_{\boldsymbol{\lambda}^0}\right)\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\left(\mathbf{I}-\mathcal{P}_{\widehat{\boldsymbol{f}}}\right)e'\right)\\ &= \mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right) - \mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right)\\ &\quad - \mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right) + \mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right),\end{aligned}$$

and, similarly to (2.7.31),

$$\begin{aligned}
\left|\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right)\right| &= \left|\sum_{k=1}^{K}\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]e'\right)\right| \\
&\leq \sum_{k=1}^{K} 2\left\|\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]e'\right\|_2 \\
&\leq 2\left\|e\right\|_2 \times \sum_{k=1}^{K}\left\|\left[\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]\right\|_2 \\
&\leq 2\left\|e\right\|_2 \times \sum_{k=1}^{K}\left\|\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\right\|_2 \times \left\|\mathbf{X}_k\right\|_2 \\
&\leq 2\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}. \\
\left|\mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right| &\leq 2\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}. \\
\left|\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right| &\leq 2\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}.
\end{aligned}$$

Then, we obtain:

$$\begin{aligned}
&\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{M}_{\widehat{\boldsymbol{f}}}e'\right) \\
&\quad\geq \mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right) - 6\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}. \qquad (2.7.33)
\end{aligned}$$

Plugging (2.7.32) and (2.7.33) in (2.7.30), we obtain:

$$\begin{aligned}
\mathrm{Tr}\left(ee'\right) \geq\ &\mathrm{Tr}\left(ee'\right) + \mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right)\mathcal{M}_{\widehat{\boldsymbol{f}}}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right)'\right) \\
&+ 2\,\mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right) \\
&- 6\left\|e\right\|_2^2 - 12\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}.
\end{aligned} \qquad (2.7.34)$$

Under Assumption 2.7.1, Lemma S.6 of Fernández-Val and Weidner (2016) holds and implies that $\|e\|_2 = O_P(N^{5/8})$. Moreover, $\sqrt{\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2} = O_P(\sqrt{NT}) = O_P(N)$.

Now, denote by $\tilde{\boldsymbol{\beta}}$ the solutions of the MLE with $\alpha_i = \alpha_i^0$ and $\xi_t = \xi_t^0$ for $i = 1,\ldots,N$ and $t = 1,\ldots,T$. We prove the following lemma.

Lemma 2.7.7 *Suppose that Assumption 2.7.1 holds. Then,* $\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} = O_P\left(\sqrt{\frac{\ln N}{N}}\right)$.

Proof: The first-order condition of the MLE with respect to $\tilde{\boldsymbol{\beta}}$ is: for $k = 1, \ldots, K$ and $i = 1, \ldots, N$,

$$\sum_{t=1}^{T} \partial_z \ell_{it}(\tilde{z}_{it}) x_{it}^{(k)} = 0. \tag{2.7.35}$$

Then, using the first-order Taylor expansion at $\beta_i^{(k)0}$, we obtain:

$$\begin{aligned} 0 &= \sum_{t=1}^{T} \partial_z \ell_{it}(z_{it}^0) x_{it}^{(k)} + \sum_{t=1}^{T} \partial_{z^2} \ell_{it}(z_{it}^*)(x_{it}^{(k)})^2 \times (\tilde{\beta}_i^{(k)} - \beta_i^{(k)0}) \\ \Longrightarrow \tilde{\beta}_i^{(k)} - \beta_i^{(k)0} &= \frac{\frac{1}{T}\sum_{t=1}^{T} \partial_z \ell_{it}(z_{it}^0) x_{it}^{(k)}}{\frac{1}{T}\sum_{t=1}^{T} [-\partial_{z^2} \ell_{it}(z_{it}^*)](x_{it}^{(k)})^2}, \end{aligned}$$

where z_{it}^* is between z_{it}^0 and $\tilde{z}_{it}$. Since Assumption 2.7.1(iv) ensures that

$$\frac{1}{T}\sum_{t=1}^{T} [-\partial_{z^2} \ell_{it}(z_{it}^*)](x_{it}^{(k)})^2 \geq b_{\min} C_2,$$

we have

$$\left|\tilde{\beta}_i^{(k)} - \beta_i^{(k)0}\right| \leq \left|\frac{1}{T}\sum_{t=1}^{T} \partial_z \ell_{it}(z_{it}^0) x_{it}^{(k)}\right| / b_{\min} C_2 =: \left|S_{k,i,T}\right| / b_{\min} C_2.$$

Without loss of generality, suppose $K = 1$ and we drop the notation k in the following. Note that given $\{x_{it}\}_{i=1,\ldots,N;t=1,\ldots,T}, \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0, \boldsymbol{\xi}^0$, $S_{i,T}$ is a sum of independent random variables. Let $M := \max_{i,t,k} \inf\{m > 0 : \mathbb{E}[\exp(|\partial_z \ell_{it}(z_{it}^0) x_{it}| / m)] \leq 2\}$. Under Assumption 2.7.1(iii) and using Bernstein inequality (see, e.g., Corollary 2.8.3 in Vershynin, 2019), we obtain that there exists $C_4 > 0$ such that for N, T sufficiently large:

$$\begin{aligned} \mathbb{P}\left(\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_{\max} \leq A\sqrt{\frac{\ln N}{N}}\right) &= \prod_{i=1}^{N} \mathbb{P}\left(\left|\tilde{\beta}_i - \beta_i^0\right| \leq A\sqrt{\frac{\ln N}{N}}\right) \\ &\geq \prod_{i=1}^{N} \mathbb{P}\left(|S_{i,T}| \leq A b_{\min} C_2 \sqrt{\frac{\ln N}{N}}\right) \\ &= \prod_{i=1}^{N} \left(1 - \mathbb{P}\left(|S_{i,T}| > A b_{\min} C_2 \sqrt{\frac{\ln N}{N}}\right)\right) \\ &\geq \left(1 - 2\exp\left\{-A^2 B \times \ln N\right\}\right)^N \\ &= \left(1 - \frac{2}{N^{A^2 B}}\right)^N, \end{aligned}$$

where $B = \frac{C_4 b_{\min}^2 C_2^2}{\kappa M^2}$. As a result, for $A > \frac{\sqrt{\kappa} M}{\sqrt{C_4} b_{\min} C_2}$ and $N, T \to \infty$, we have:

$$\mathbb{P}\left(\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_{\max} \leq A\sqrt{\frac{\ln N}{N}}\right) \to 1.$$

and therefore $\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}=O_P\left(\sqrt{\frac{\ln N}{N}}\right)$. The proof is completed. □

Using (2.7.35), we obtain that: for $k=1,\ldots,K$ and $i=1,\ldots,N$,

$$\sum_{t=1}^{T}\partial_z\ell_{it}(\tilde{z}_{it})x_{it}^{(k)}=0 \implies \text{The diagonal elements of } \mathbf{X}_k\tilde{e}' \text{ are zeros.}$$
$$\implies \operatorname{Tr}\left(\operatorname{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\tilde{e}'\right)=0,$$

where $\tilde{e}=(\partial_z\ell_{it}(x_{it}'\tilde{\beta}_i+\alpha_i^0+\xi_t^0))_{i=1,\ldots,N;t=1,\ldots,T}$. Then,

$$\begin{aligned}
\left|\operatorname{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right)\right| &= \left|\operatorname{Tr}\left(\sum_{k=1}^{K}\left[\operatorname{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right](e-\tilde{e})'\right)\right| \\
&= \left|\operatorname{Tr}\left(\sum_{k=1}^{K}\left[\operatorname{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]\left[\int_0^1\partial_{z^2}\ell_{it}\left(x_{it}'(t\beta_i^0+(1-t)\tilde{\beta}_i)+\alpha_i^0+\xi_t^0\right)\sum_{k=1}^{K}(\beta_i^{(k)0}-\tilde{\beta}_i^{(k)})x_{it}^{(k)}dt\right]'_{i=1,\ldots,N;t=1,\ldots,T}\right)\right| \\
&\le b_{\max}\operatorname{Tr}\left(\sum_{k=1}^{K}\left[\operatorname{Diag}\left(\left|\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right|\right)|\mathbf{X}_k|\right]\sum_{k=1}^{K}\left[\operatorname{Diag}\left(\left|\tilde{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right|\right)|\mathbf{X}_k|\right]'\right) \\
&\le Kb_{\max}\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}\times\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}\times\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2
\end{aligned} \tag{2.7.36}$$

Plugging (2.7.36) in (2.7.34), we obtain:

$$\begin{aligned}
&6\|e\|_2^2+2\sqrt{K\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2}\left(6\|e\|_2+b_{\max}\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}\sqrt{K\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2}\right)\times\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \\
&\quad\ge \operatorname{Tr}\left(\mathcal{M}_{(\boldsymbol{\alpha}^0,\mathbf{1}_N)}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right)\mathcal{M}_{(\mathbf{1}_T,\widehat{\boldsymbol{\xi}})}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right)'\right).
\end{aligned}$$

Using Assumption 2.7.1(v), we obtain that

$$\begin{aligned}
&\frac{1}{NT}\left[6\|e\|_2^2+2\sqrt{K\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2}\left(6\|e\|_2+b_{\max}\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}\sqrt{K\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2}\right)\times\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}\right] \\
&\quad\ge c\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}^2.
\end{aligned}$$

with probability one. Then, this inequality, together with $\|e\|_2=O_P(N^{5/8})$, $\sqrt{\sum_{k=1}^{K}\|\mathbf{X}_k\|_{\mathrm{F}}^2}=O_P(N)$, and Lemma 2.7.7, implies that $\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max}=O_P(N^{-3/8})$.

2.7.11 Monte Carlo Experiments: Additional Tables and Details

TABLE 2.7: Numerical Convergence – Logit Model with Homogeneous Slopes ($N=5000$, $T=30$)

	FPMLE^{++}			FPMLE			MLE
	RMSE to MLE		CPU time	RMSE to MLE		CPU time	CPU time
#iter	$\widehat{\beta}^{++}$	$\widehat{\delta}^{++}$		$\widehat{\beta}$	$\widehat{\delta}$		
1	0.21091	0.03081	0.06655	0.04698	0.00587	2.16765	1131.27882
3	0.00204	0.00018	0.11279	0.00454	0.00045	5.16679	
20	0.00001	0.00000	0.65884	0.00451	0.00045	9.19747	

Notes: Each row reports the results obtained after **#iter** iterations (for FPMLE and FPMLE^{++}) and based on 50 replications for DGP (i).

Table 2.8 summarizes the bias of $\widehat{\beta}_i$ (estimated by FPMLE and FPMLE^{++}) and its numerical convergence to the MLE estimator.

TABLE 2.8: Numerical Convergence – Poisson Model with Heterogeneous Slopes ($N = T = 200$)

		FPMLE^{++}		**FPMLE**	
DGP	**#iter**	**Bias $\widehat{\beta}$**	**RMSE to MLE $\widehat{\beta}$**	**Bias $\widehat{\beta}$**	**RMSE to MLE $\widehat{\beta}$**
i.	1	-0.30899	0.51512	0.01421	0.13466
	3	-0.13506	0.23583	0.02586	0.07470
	20	-0.00027	0.10394	0.00066	0.00547
ii.	1	-0.37449	0.50871	0.02664	0.16904
	3	-0.11055	0.18596	0.01994	0.09795
	20	0.00536	0.04741	0.00322	0.00872
iii.	1.	-0.49078	0.76451	0.12356	0.16154
	3.	-0.28560	0.56801	0.20954	0.22794
	20.	-0.14200	0.39700	0.01048	0.01270
iv.	1	-0.59094	0.87637	0.08003	0.14778
	3	-0.29487	0.50883	0.17232	0.18473
	20	-0.03348	0.24580	0.00742	0.00829

Notes: The biases are computed as $\frac{1}{1000}\sum_{r=1}^{50}\sum_{i=1}^{200}\widehat{\beta}_i^{(r)} - \beta_{0,i}^{(r)}$.

Split-sample Jackknife Bootstrap Procedure

For any finite non-empty set of indices I and $n \in I$, let $I_{:n}$ (resp. $I_{n:}$) denote indices up to the nth indice (resp. after the $n+1$th) in I. We describe the procedure for the case of a scalar β_i. The extension to the multidimensional case is straightforward. Denote by B the number of bootstrap samples.

Bootstrap Percentile CI's:

1. For $b \in \{1, \dots, B\}$:

   (a) Draw N^* units from $\{1, \dots, N\}$ with replacement, sort them and label them with indices $\mathcal{I}^{(b)}$ such that $\left|\mathcal{I}^{(b)}\right| = N^*$.

   (b) Compute full-sample FPMLE^{++} estimates using $\mathcal{I}^{(b)}$:

   $$\left\{\widehat{\beta}^{(b)}_{i,\text{fs}} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\alpha}^{(b)}_{i,\text{fs}} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\xi}^{(b)}_{t,\text{fs}} : t = 1, \dots, T\right\}.$$

   (c) Compute half-sample FPMLE^{++} estimates using only units in $\mathcal{I}^{(b)}_{:\lfloor N^*/2 \rfloor}$:

   $$\left\{\widehat{\beta}^{(b)}_{i,1\text{N}^*} : i \in \mathcal{I}^{(b)}_{:\lfloor N^*/2 \rfloor}\right\}, \quad \left\{\widehat{\alpha}^{(b)}_{i,1\text{N}^*} : i \in \mathcal{I}^{(b)}_{:\lfloor N^*/2 \rfloor}\right\}, \quad \left\{\widehat{\xi}^{(b)}_{t,1\text{N}^*} : t = 1, \dots, T\right\}.$$

   Compute half-sample FPMLE^{++} estimates using only units in $\mathcal{I}^{(b)}_{\lfloor N^*/2 \rfloor:}$:

   $$\left\{\widehat{\beta}^{(b)}_{i,2\text{N}^*} : i \in \mathcal{I}^{(b)}_{\lfloor N^*/2 \rfloor:}\right\}, \quad \left\{\widehat{\alpha}^{(b)}_{i,2\text{N}^*} : i \in \mathcal{I}^{(b)}_{\lfloor N^*/2 \rfloor:}\right\}, \quad \left\{\widehat{\xi}^{(b)}_{t,2\text{N}^*} : t = 1, \dots, T\right\}.$$

   (d) Compute half-sample FPMLE^{++} estimates using only time periods in $\{1, \dots, \lfloor T/2 \rfloor\}$:

   $$\left\{\widehat{\beta}^{(b)}_{i,1\text{T}} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\alpha}^{(b)}_{i,1\text{T}} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\xi}^{(b)}_{t,1\text{T}} : t = 1, \dots, \lfloor T/2 \rfloor\right\}.$$

   Compute half-sample FPMLE^{++} estimates using only time periods in $\{\lfloor T/2 \rfloor + 1, \dots, T\}$:

   $$\left\{\widehat{\beta}^{(b)}_{i,2\text{T}} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\alpha}^{(b)}_{i,2\text{T}} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\xi}^{(b)}_{t,2\text{T}} : t = \lfloor T/2 \rfloor + 1, \dots, T\right\}.$$

(e) Let

$$\widehat{\mu}^{(b)} := 3\left(\frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\widehat{\beta}^{(b)}_{i,\mathrm{fs}}\right) - \frac{1}{2}\left(\frac{1}{\lfloor N^*/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}_{:\lfloor N^*/2\rfloor}}\widehat{\beta}^{(b)}_{i,1\mathrm{N}^*} + \frac{1}{N^*-\lfloor N^*/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}_{\lfloor N^*/2\rfloor:}}\widehat{\beta}^{(b)}_{i,2\mathrm{N}^*}\right)$$
$$- \frac{1}{2}\left(\frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\widehat{\beta}^{(b)}_{i,1\mathrm{T}} + \frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\widehat{\beta}^{(b)}_{i,2\mathrm{T}}\right),$$

$$\widehat{\sigma}^{(b)} = 3\sqrt{\frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\left(\widehat{\beta}^{(b)}_{i,\mathrm{fs}} - \frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\widehat{\beta}^{(b)}_{i,\mathrm{fs}}\right)^2}$$
$$- \frac{1}{2}\left(\sqrt{\frac{1}{\lfloor N^*/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}_{:\lfloor N^*/2\rfloor}}\left(\widehat{\beta}^{(b)}_{i,1\mathrm{N}^*} - \frac{1}{\lfloor N^*/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}_{:\lfloor N^*/2\rfloor}}\widehat{\beta}^{(b)}_{i,1\mathrm{N}^*}\right)^2}\right.$$
$$\left. + \sqrt{\frac{1}{N^*-\lfloor N^*/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}_{\lfloor N^*/2\rfloor:}}\left(\widehat{\beta}^{(b)}_{i,2\mathrm{N}^*} - \frac{1}{N^*-\lfloor N^*/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}_{\lfloor N^*/2\rfloor:}}\widehat{\beta}^{(b)}_{i,2\mathrm{N}^*}\right)^2}\right)$$
$$- \frac{1}{2}\left(\sqrt{\frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\left(\widehat{\beta}^{(b)}_{i,1\mathrm{T}} - \frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\widehat{\beta}^{(b)}_{i,1\mathrm{T}}\right)^2}\right.$$
$$\left. + \sqrt{\frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\left(\widehat{\beta}^{(b)}_{i,2\mathrm{T}} - \frac{1}{N^*}\sum_{i\in\mathcal{I}^{(b)}}\widehat{\beta}^{(b)}_{i,2\mathrm{T}}\right)^2}\right),$$

$$\widehat{\mathrm{APE}}^{(b)} = 3\left(\frac{1}{N^*T}\sum_{i\in\mathcal{I}^{(b)}}\sum_{t=1}^{T}\sum_{y\in\mathcal{Y}} y\widehat{\beta}^{(b)}_{i,\mathrm{fs}}g'(y; x_{it}\widehat{\beta}^{(b)}_{i,\mathrm{fs}} + \widehat{\alpha}^{(b)}_{i,\mathrm{fs}} + \widehat{\xi}^{(b)}_{t,\mathrm{fs}})\right)$$
$$- \frac{1}{2}\left(\frac{1}{\lfloor N^*/2\rfloor T}\sum_{i\in\mathcal{I}^{(b)}_{:\lfloor N^*/2\rfloor}}\sum_{t=1}^{T}\sum_{y\in\mathcal{Y}} y\widehat{\beta}^{(b)}_{i,1\mathrm{N}^*}g'(y; x_{it}\widehat{\beta}^{(b)}_{i,1\mathrm{N}^*} + \widehat{\alpha}^{(b)}_{i,1\mathrm{N}^*} + \widehat{\xi}^{(b)}_{t,1\mathrm{N}^*})\right.$$
$$\left. + \frac{1}{(N^*-\lfloor N^*/2\rfloor)T}\sum_{i\in\mathcal{I}^{(b)}_{\lfloor N^*/2\rfloor:}}\sum_{t=1}^{T}\sum_{y\in\mathcal{Y}} y\widehat{\beta}^{(b)}_{i,2\mathrm{N}^*}g'(y; x_{it}\widehat{\beta}^{(b)}_{i,2\mathrm{N}^*} + \widehat{\alpha}^{(b)}_{i,2\mathrm{N}^*} + \widehat{\xi}^{(b)}_{t,2\mathrm{N}^*})\right)$$
$$- \frac{1}{2}\left(\frac{1}{N^*\lfloor T/2\rfloor}\sum_{i\in\mathcal{I}^{(b)}}\sum_{t=1}^{\lfloor T/2\rfloor}\sum_{y\in\mathcal{Y}} y\widehat{\beta}^{(b)}_{i,1\mathrm{T}}g'(y; x_{it}\widehat{\beta}^{(b)}_{i,1\mathrm{T}} + \widehat{\alpha}^{(b)}_{i,1\mathrm{T}} + \widehat{\xi}^{(b)}_{t,1\mathrm{T}})\right.$$
$$\left. + \frac{1}{N^*(T-\lfloor T/2\rfloor)}\sum_{i\in\mathcal{I}^{(b)}}\sum_{t=\lfloor T/2\rfloor+1}^{T}\sum_{y\in\mathcal{Y}} y\widehat{\beta}^{(b)}_{i,2\mathrm{T}}g'(y; x_{it}\widehat{\beta}^{(b)}_{i,2\mathrm{T}} + \widehat{\alpha}^{(b)}_{i,2\mathrm{T}} + \widehat{\xi}^{(b)}_{t,2\mathrm{T}})\right),$$

and $\widehat{\mu} := (\widehat{\mu}^{(1)}, \ldots, \widehat{\mu}^{(B)})$, $\widehat{\sigma} := (\widehat{\sigma}^{(1)}, \ldots, \widehat{\sigma}^{(B)})$, and $\widehat{\mathrm{APE}} := (\widehat{\mathrm{APE}}^{(1)}, \ldots, \widehat{\mathrm{APE}}^{(B)})$.

2. For $\alpha \in (0,1)$, build CI's

$$\left[q_{\alpha/2}(\widehat{\mu}), q_{1-\alpha/2}(\widehat{\mu})\right], \quad \left[q_{\alpha/2}(\widehat{\sigma}), q_{1-\alpha/2}(\widehat{\sigma})\right], \quad \left[q_{\alpha/2}(\widehat{\mathrm{APE}}), q_{1-\alpha/2}(\widehat{\mathrm{APE}})\right],$$

where $q_u(\mathbf{X})$ is the uth empirical quantile of the sample $\mathbf{X}$.

Chapter 3

A Simple and Computationally Trivial Estimator for Grouped Fixed Effects Models

Il peut paraître étonnant que les pensées profondes se rencontrent plutôt dans les écrits des poètes que dans ceux des philosophes. La raison en est que les poètes ont écrit sous l'empire de l'enthousiasme et de la force de l'imagination. Il y a en nous des semences de science, comme en un silex ; les philosophes les extraient par raison ; les poètes les arrachent par imagination : elles brillent alors davantage.

René Descartes, *Discours de la méthode*

Abstract: This chapter provides a new fixed effects estimator for linear panel data models with clustered time patterns of unobserved heterogeneity. The method combines a preliminary consistent estimate for the slope coefficient (e.g., using smooth and convex nuclear norm regularization) with a pairwise differencing argument that takes at most $O(N^3T)$ elementary operations to cluster cross-sectional units. Asymptotic guarantees are established in a framework where T can grow at any power of N, as both N and T diverge to infinity. Unlike most existing approaches, the new estimator (i) is computationally straightforward, (ii) does not require a known upper bound on the number of groups, and (iii) consistently estimates the number of groups. As most existing approaches, it (iv) correctly classifies units into groups with probability tending to one, (v) is asymptotically equivalent to the infeasible least squares estimator which controls for the true group indicators, (vi) and is asymptotically normal unbiased at parametric rates.

3.1 Introduction

Consider the grouped fixed effects model:

$$y_{it} = x_{it}'\beta + \alpha_{g_i t} + v_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T, \tag{3.1.1}$$

where i denotes cross-sectional units, t denotes time periods, $y_{it} \in \mathbb{R}$ is a dependent observed variable, and $x_{it} \in \mathbb{R}^p$ is a vector of observed covariates uncorrelated with the zero-mean random variable $v_{it} \in \mathbb{R}$, but potentially correlated with the group-specific time-varying unobservable $\alpha_{g_i t} \in \mathbb{R}$. The common vector $\beta \in \mathbb{R}^p$, the number of groups $G \in \mathbb{N}^*$, the group membership variables $g_i \in \{1, \dots, G\}$, and the vectors of group-specific time effects $(\alpha_{1t}, \dots, \alpha_{Gt})' \in \mathbb{R}^G$ are unrestricted and to be estimated.

This special case of interactive fixed effects models (e.g., Bai, 2009) offers a flexible parcimonious way to model time-varying unobserved heterogeneity and cross-sectional correlations.[1] It was first proposed in Bonhomme and Manresa (2015). In the following, N and T are assumed relatively large, while p and G relatively small. Interest lies in the high-dimensional parameter $\theta = (G, g_1, \dots, g_N, \alpha_{11}, \dots, \alpha_{GT}, \beta')'$.

To the best of my knowledge, the panel data literature does not mention the existence of an estimator for θ that, in an asymptotic framework where T can grow at any power of N, (i) is computationally feasible, (ii) does not require a known upper bound on the number of groups, (iii) consistently estimates the number of groups, (iv) uniformly correctly classifies units into groups with probability approaching one, (v) is asymptotically equivalent to the oracle OLS regression that controls for the true group indicators, (vi) and is asymptotically normal unbiased at the $\sqrt{NT}$ and $\sqrt{N}$ parametric rates for β and α_{gt}. These properties are useful to make inference on θ when T is only moderately large with respect to N (e.g., microeconometric datasets).

First, I introduce a new and simple two-step estimator $\widehat{\theta}$ that enjoys properties (i)-(ii). The first step is a clustering step. Given an off-the-shelve preliminary consistent estimate for β, I apply a "triad pairwise differencing" tranformation to the residualized outcome in order to compute pairwise distances betwen units. Close units are grouped together according to the outcome of $N(N-1)/2$ tests, rejecting $H_{0ij} : g_i = g_j$ if and only if the pairwise distance between i and j falls above some regularization positive threshold c_{NT}. Specifically, units with identical test outcomes are grouped together. The second step is a pooled OLS regression of the outcome variable on the covariates and the interactions of time and estimated group dummies. The interactive fixed effects literature provides many candidates for computationally simple preliminary consistent estimates of β (e.g., nuclear norm, square-root lasso), which are discussed more in details in Section 3.3.3.

[1] Note $\alpha_{g_i t} = \lambda_i' F_t$ with $\lambda_i' = (c\mathbf{1}\{g_i = 1\}, \dots, c\mathbf{1}\{g_i = G\})$, $F_t' = (\alpha_{1t}/c, \dots, \alpha_{Gt}/c)$, $c \neq 0$. Thus, the re-scaled vector of factor loadings, λ_i/c, lies in the finite set of vertices of the unit simplex of $\mathbb{R}^G$. Reciprocally, if $\left|\left\{\tilde{\lambda}_1, \dots, \tilde{\lambda}_N\right\}\right| = \tilde{G}$, there exists $(\tilde{g}_1, \dots, \tilde{g}_N, \tilde{\alpha}_{11}, \dots, \tilde{\alpha}_{\tilde{G}T})'$ such that $\tilde{g}_i \in \left\{1, \dots, \tilde{G}\right\}$ and $\tilde{\lambda}_i' \tilde{F}_t = \tilde{\alpha}_{\tilde{g}_i t}$.

Second, I derive sufficient conditions for $\widehat{\theta}$ to enjoy properties (iii)-(vi). To the best of my knowledge, this work is the first to propose an estimator for model (3.1.1) with properties (i)-(vi). As a byproduct, it is also the first to show property (iii) for an estimator that also verifies property (ii) in a specific interactive fixed effects model. This property stands in sharp contrast with the general case, in which a known upper bound is generally required (see, e.g., Bai, 2009; Bai and Ng, 2002, 2019). I do not address the question of conducting inference on groups. Dzemski and Okui (2018) provides pointwise valid inference methods given a preliminary estimator $\widehat{\beta}$ is available. One can use the estimator proposed here. Overall, our results show that efficient estimation of θ is possible under the maintained asymptotic framework even if absolutely no information is available regarding the number of groups. Moreover, they suggest that the first-step estimation of the latent clustering structure neither affects the bias nor the variance of our estimator compared to the oracle OLS regression.

There is a vast literature on estimating interactive fixed effects panel models (see, e.g., Bai, 2009; Bonhomme et al., 2022; Higgins, 2022; Moon and Weidner, 2015, 2017; Pesaran, 2006). Traditional leading methods rely on non-convex least squares and Principal Component Analysis, which violate property (i) and are not known to verify the other properties. Convex nuclear norm regularization methods proposed in Moon and Weidner (2019) and Chernozhukov et al. (2019) are not known to verify properties (iv)-(vi). They can serve as preliminary consistent estimators in the present approach. Bonhomme and Manresa (2015)'s grouped fixed-effects (GFE) estimator, an extension of k-means clustering to handle covariates, solves a NP-hard optimization problem. Algorithms that provide fast solutions may not converge to its true value. The same concern applies to extensions to grouped factor models (e.g., Ando and Bai, 2017, 2022) and other non-convex estimators (e.g., Su et al., 2016). The GFE estimator does not verify properties (i)-(iii). In contrast, the inferential theory developed here is valid for a computationally simple estimator which substitutes $G_{\max}$ with a pairwise distance regularization parameter. Because inference is on a true population parameter θ, it also contrasts with Pollard (1981, 1982) which provides asymptotic theory for the solution to the population k-means sum of squares problem in the cross-sectional case, i.e., only for a pseudo-true value. Chetverikov and Manresa (2021)'s spectral and post-spectral estimators enjoy properties (i) and (iv)-(vi) by imposing a grouped factor structure on the covariates and assuming that "G can be consistently estimated". The paper does not explicitely provide a mean to do so under the maintained asymptotic framework. To the best of my knowledge, these estimators are not known to verify properties (ii)-(iii). Lewis et al. (2023) propose an approximating procedure based on the GMM framework in the general case of grouped slope coefficients, but they do not derive any asymptotic equivalence result to the oracle estimator.

While the proposed clustering method seems to be new to the literature, the pairwise distance measure used in the clustering step has already been employed in the mathematical statistics and econometric literature to study topological properties

of the graphon (e.g., Auerbach, 2022; Lovász, 2012; Zeleneev, 2020; Zhang et al., 2017). More generally, dyad, triad, or tetrad comparisons have proven useful in a variety of other econometric contexts (see, e.g., Charbonneau, 2017; Graham, 2017; Honoré and Powell, 1994; Jochmans, 2017). Albeit close in spirit, the procedure is different from the binary segmentation algorithm developed in Wang and Su (2021), or the pairwise comparisons method proposed in Krasnokutskaya et al. (2022). Some papers rely on spectral clustering (see, e.g., Brownlees et al., 2022; Chetverikov and Manresa, 2021; Ng et al., 2002; von Luxburg, 2007; Yu et al., 2022).

The operational research clustering literature is old and mature, and other agglomerative clustering approaches could be adapted (e.g., DBSCAN, HDBSCAN). The method retained here is one for which asymptotic properties are relatively straightforward to establish under transparent conditions useful in economic applications. Also, this work can be seen as the first application of an agglomerative clustering methods to the econometric panel data model (3.1.1).

Section 3.2 introduces the two-step estimator. Section 3.3 presents large sample properties, including uniform consistency for the grouping structure and asymptotic normality at parametric rates. Section 3.4 contains a brief discussion and concludes. All proofs are in the Appendix.

3.2 A Two-Step Estimator

The goal is to estimate the parameter $\theta = (G, g_1, \ldots, g_N, \alpha_{11}, \ldots, \alpha_{GT}, \beta')' \in \Theta$, where $\Theta = \bigcup_{g=1}^{+\infty} \Theta_g$ and $\Theta_g = \{g\} \times \{1, \ldots, g\}^N \times \mathcal{A}^{gT} \times \mathcal{B}$ for some $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{A} \subset \mathbb{R}$. A simple and computationally trivial estimator $\widehat{\theta} = (\widehat{G}, \widehat{g}_1, \ldots, \widehat{g}_N, \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{\widehat{G}T}, \widehat{\beta}')' \in \Theta$ can be obtained in two steps. Let $c_{NT} \in (0, \infty)$, and $\widehat{\beta}^1$ be a preliminary consistent estimator for β.[2]

1. Clustering Step:

   1.a. Compute for all $(i,t) \in \{1, \ldots, N\} \times \{1, \ldots, T\}$:

   $$\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\beta}^1.$$

   1.b. Compute for all $(i,j) \in \{1, \ldots, N\}^2$:

   $$\widehat{d}_{\infty}^2(i,j) = \max_{k \in \{1,\ldots,N\} \setminus \{i,j\}} \left| \frac{1}{T} \sum_{t=1}^{T} (\widehat{v}_{it} - \widehat{v}_{jt}) \widehat{v}_{kt} \right|.$$

   1.c. Compute for all $(i,j) \in \{1, \ldots, N\}^2$:

   $$\widehat{W}_{ij} = \mathbf{1}\Big\{ \widehat{d}_{\infty}^2(i,j) \leq c_{NT} \Big\}.$$

[2] Examples are given in Section 3.3.3.

1.d. Set $k = 1$.
Set $\widehat{g}_1 = 1$ and $\widehat{C}_1 = \left\{ i \in \{1, \ldots, N\} : \widehat{W}_{ij} = \widehat{W}_{1j} \quad \forall j \in \{1, \ldots, N\} \right\}$.
For all $i \in \widehat{C}_1$: set $\widehat{g}_i = 1$.
Set `stop`=False.
While not `stop`:

- Set $i_k^* = \inf \left\{ i \in \{1, \ldots, N\} : i \notin \cup_{\ell=1}^{k} \widehat{C}_\ell \right\}$.
- If $i_k^* < \infty$:
  - Set $\widehat{C}_{k+1} = \left\{ i \in \{1, \ldots, N\} : \widehat{W}_{ij} = \widehat{W}_{i_k^* j} \quad \forall j \in \{1, \ldots, N\} \right\}$.
  - For all $i \in \widehat{C}_{k+1}$: set $\widehat{g}_i = k + 1$.
  - Set $k = k + 1$.
- Else:
  - Set `stop`=True.

Set $\widehat{G} = |\{\widehat{g}_1, \ldots, \widehat{g}_N\}|$.

2. Projection Step:
Compute:

$$\left(\widehat{\beta}', \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{\widehat{G}T}\right) \in \underset{\left(\beta', \alpha_{11}, \ldots, \alpha_{\widehat{G}T}\right) \in \mathcal{B} \times \mathcal{A}^{\widehat{G}T}}{\arg\max} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\beta - \alpha_{\widehat{g}_i t} \right)^2 .$$

The estimation procedure has two steps. In the clustering step, a binary thresholding operator is applied coordinate-wise to the dissimilarity matrix $\widehat{D} := (\widehat{d}_\infty^2(i,j))_{(i,j) \in \{1,\ldots,N\}^2}$ to obtain the binary adjacency matrix $\widehat{W}$. The tuning parameter c_{NT} is the largest possible value for $\widehat{d}_\infty^2(i,j)$ that still yields an estimated undirected edge between units i and j. Then, estimated groups are formed sequentially by regrouping units with identical sets of edges: start by gathering in Group 1 units that share the same set of estimated edges as unit 1, then gathering in Group 2 units that share the same set of estimated edges as the unit with smallest index not already in Group 1, etc. By construction, this estimation procedure always assigns all units to $\widehat{G} \in \{1, \ldots, N\}$ non-empty groups which contain units with the same set of estimated edges. Intuitively, if $c_{NT} \to 0$, $\widehat{G} \to N$; if $c_{NT} \to \infty$, $\widehat{G} \to 1$. In the projection step, a pooled OLS regression of y_{it} on x_{it} and the interactions of groups and time dummies (in the special case where $\mathcal{B} = \mathbb{R}^p$ and $\mathcal{A} = \mathbb{R}$) delivers $\left(\widehat{\beta}', \widehat{\alpha}_{11}, \ldots, \widehat{\alpha}_{\widehat{G}T}\right)'$.

In Section 3.3, we provide theoretical guidance for the choice of c_{NT}. In finite sample, it can be chosen by cross-validation (see Section 3.3.4).

Remark 1: Consider a finite sample of fixed dimensions N and T. If all random variables except group memberships are continuous, then to the extreme where $c_{NT} \to 0$, $\widehat{G} \to N$ and each group contains a single unit. To the extreme where $c_{NT} \to +\infty$, $\widehat{G} \to 1$ and a single group contains all units. Given the low CPU time of the method, an entire regularization path can be reported by the researcher by making

c_{NT} vary between these two regimes. In particular, the clustering step can be made all vectorized, greatly reducing computational burden compared to running loops.[3] Also, it is possible to improve the finite sample performance by re-running the first step with $\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\beta}$ in place of $\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\beta}^1$ to obtain new $\widehat{g}_1, \ldots, \widehat{g}_N$, and then re-running the second step and iterating again until some convergence criterion is achieved. The asymptotic results will hold for any of the subsequent iterates.

Remark 2: When unobserved heterogeneity is assumed to be time-constant, the $O(N^3T)$ computation cost can be reduced to $O(N^2T)$ and the preliminary estimator can be replaced with any standard differencing fixed effects estimator such as Arellano and Bond (1991). See also Wooldridge (2010).

Remark 3: The intuition for the estimator is as follows (see also p.14 in Zeleneev, 2020). Asymptotically, $\widehat{v}_{it} \approx \alpha_{g_it}$ so that $g_i = g_j$ implies, "uniformly" over k,

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt} \approx 0 \\ \implies &\max_{k\in\{1,\ldots,N\}\setminus\{i,j\}} \left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt}\right| \approx 0. \end{aligned}$$

Reciprocally, if

$$\max_{k\in\{1,\ldots,N\}\setminus\{i,j\}} \left|\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt}\right| \approx 0, \tag{3.2.1}$$

then necessarily $g_i = g_j$. To see it, note that if $g_i \neq g_j$, and provided each group has at least 2 units asymptotically (which is weak), then there exist $k^*, l^* \in \{1, \ldots, N\} \setminus \{i, j\}$ such that $g_{k^*} = g_i$ and $g_{l^*} = g_j$. Equation (3.2.1) implies in turn

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{k^*t} \approx 0, \tag{3.2.2}$$

$$\frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{l^*t} \approx 0. \tag{3.2.3}$$

Differencing (3.2.2)-(3.2.3) yields

$$\frac{1}{T}\sum_{t=1}^{T}(\alpha_{g_it} - \alpha_{g_jt})^2 \approx 0,$$

a contradiction if groups are well separated, e.g., if for all $(g, \widetilde{g}) \in \{1, \ldots, G\}^2$, $g \neq \widetilde{g}$, there exists $c_{g,\widetilde{g}} > 0$ such that

$$\frac{1}{T}\sum_{t=1}^{T}(\alpha_{gt} - \alpha_{\widetilde{g}t})^2 \geq c_{g,\widetilde{g}} > 0.$$

[3] MATLAB code is provided on the author's website: https://martinmugnier.github.io/research. A small Monte Carlo exercise and an empirical illustration are presented in sections S.3-4 of the Supplemental Material.

The next section formalizes the identification result.

3.3 Large Sample Properties

Consider the following data generating process:

$$y_{it} = x_{it}'\beta^0 + \alpha^0_{g_i^0 t} + v_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T, \tag{3.3.1}$$

where $g_i^0 \in \{1, \dots, G^0\}$ denotes group membership, and where the 0 superscripts refer to true parameter values. I consider asymptotic sequences (N, T) where N and T diverge jointly to infinity. I assume for now that the number of groups G^0 is fixed (relative to N, T) but unknown, and I defer the discussion on the case of an increasing sequence $G^0 = G^0_{NT}$ to the Supplemental Material S.1.[4]

For any set $\mathcal{I} \subset \mathbb{N}^*$, all $k \in \mathbb{N}^*$, let $\mathscr{P}_k(\mathcal{I})$ collects the cardinal-k subsets of $\mathcal{I}$.

3.3.1 Clustering Consistency

Consider the following assumptions.

Assumption 3.3.1 *There exists a deterministic sequence r_{NT} such that, as N and T diverge jointly to infinity, $\left\|\widehat{\beta}_1 - \beta^0\right\| = O_p(r_{NT})$ and $r_{NT} \to 0$.*

Assumption 3.3.2 *There exists $(\nu, \kappa) \in (0, +\infty) \times (0, 1/2)$ such that, as N and T diverge jointly to infinity, $NT^{-\nu} \to 0$, $c_{NT} \to 0$, $c_{NT} r_{NT}^{-1} \to \infty$. There exists $(C, N_0, T_0) \in (0, +\infty) \times \mathbb{N} \times \mathbb{N}$ such that for all (N, T) such that $N \geq N_0$ and $T \geq T_0, c_{NT} T^{\kappa} \geq C$.*

Assumption 3.3.3 *There exist constants $a, b, c, d_1, d_2 > 0$ and a sequence $\tau(t) \leq e^{-at^{d_1}}$ such that:*

(a) $\mathcal{A}$ is a compact subset of $\mathbb{R}$.

(b) For all $(i, t) \in \{1, \dots, N\} \times \{1, \dots, T\}$: $\mathbb{P}(|v_{it}| > m) \leq e^{1-(m/b)^{d_2}}$ for all $m > 0$ and $\mathbb{E}(v_{it}) = 0$.

(c) For all $(g, \widetilde{g}) \in \{1, \dots, G^0\}^2$ such that $g \neq \widetilde{g}$: $\operatorname{plim}_{T\to\infty} T^{-1} \sum_{t=1}^T (\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})^2 = c_{g,\widetilde{g}} \geq c$.

(d) For all $(i, j, k, g, \widetilde{g}) \in \mathscr{P}_3(\{1, \dots, N\}) \times \{1, \dots, G^0\}^2$ such that $g \neq \widetilde{g}$, $\{v_{it}\}_t$, $\{(v_{it} - v_{jt})v_{kt}\}_t$, $\left\{\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right\}_t$, and $\left\{(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})v_{it}\right\}_t$ are strongly mixing processes with mixing coefficients $\tau(t)$. Moreover, $\mathbb{E}((\alpha^0_{gt} - \alpha^0_{\widetilde{g}t})v_{it}) = \mathbb{E}(v_{it}v_{jt}) = 0$.

(e) $\lim_{N\to\infty} \mathbb{P}(\min_{g\in\{1,\dots,G^0\}} \sum_{i=1}^N \mathbf{1}\{g_i^0 = g\} \geq 2) = 1$.

[4] In particular, when $\beta^0 = 0$ is known, the number of groups can increase with sample size at any rate bounded by $N/2$.

(f) There exists a constant $M > 0$ such that, as N, T tend to infinity:

$$\sup_{i\in\{1,\dots,N\}} \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 \geq M\right) = o(T^{-\delta}) \text{ for all } \delta > 0,$$

where $\|\cdot\|$ denotes the Euclidean norm.

Assumption 3.3.1 requires $\widehat{\beta}_1$ to be consistent for β^0 at a rate bounded by r_{NT}. This rate can be slow. Examples of computationally simple estimators satisfying this condition are given in Section 3.3.3. Assumption 3.3.2 allows T to grow considerably more slowly than N (when $\nu \gg 1$). It requires that the tuning parameter decreases to zero, but not too fast, at a rate bounded below by $T^{-1/2}$ and strictly slower than r_{NT}. Assumptions 3.3.3(a)-(b) and 3.3.3(d) collect standard moment, tail, and dependence conditions. They do not impose homoscedasticity, but only require uniform bounds on the unconditional variances. Assumption 3.3.3(c) requires groups to be well-separated. Assumption 3.3.3(e) allows for asymptotically negligible groups, but requires that each group has at least two members with probability approaching one. Assumption 3.3.3(f) is a slight reinforcement of Bonhomme and Manresa (2015)'s Assumption 2(e). It holds if covariates have bounded support or if they satisfy dependence and tail conditions similar to v_{it}. All results below are understood up to group relabeling.

Proposition 3.3.1 (Sup-norm classification consistency) *Let Assumptions 3.3.1-3.3.3 hold. Then, as N and T tend to infinity,*

$$\sup_{i\in\{1,\dots,N\}} \left|\widehat{g}_i - g_i^0\right| = o_p(1), \tag{3.3.2}$$

and

$$\widehat{G} - G^0 = o_p(1). \tag{3.3.3}$$

3.3.2 Asymptotic Distribution

The following assumption is useful to establish the asymptotic distribution of $\widehat{\beta}$ and $\widehat{\alpha}_{gt}$.

Assumption 3.3.4

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

(b) For all $(g,t) \in \{1,\dots,G^0\} \times \{1,\dots,T\}$:

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

(c) For all $(g,t) \in \{1,\dots,G^0\} \times \{1,\dots,T\}$*: as* N *and* T *tend to infinity,*

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} v_{it} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \omega_{gt}\right),$$

where $\xrightarrow{d}$ *denotes convergence in distribution.*

(d) For all $(i,j,t) \in \{1,\dots,N\}^2 \times \{1,\dots,T\}$*:* $\mathbb{E}(x_{jt} v_{it}) = 0$*.*

(e) There exist positive definite matrices Σ_β *and* Ω_β *such that*

$$\Sigma_\beta = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \overline{x}_{g_i^0 t})(x_{it} - \overline{x}_{g_i^0 t})',$$
$$\Omega_\beta = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}\left[v_{it} v_{js} (x_{it} - \overline{x}_{g_i^0 t})(x_{js} - \overline{x}_{g_j^0 s})' \right],$$

where $\overline{x}_{gt} := \left(\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}\right)^{-1} \sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} x_{it}$*.*

(f) As N *and* T *tend to infinity:* $\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \overline{x}_{g_i^0 t}) v_{it} \xrightarrow{d} \mathcal{N}(0, \Omega_\beta)$*.*

Corollary 3.3.2 (Asymptotic distribution) *Let Assumptions 3.3.1-3.3.4 hold. Then, as* N *and* T *tend to infinity,*

$$\sqrt{NT}(\widehat{\beta} - \beta^0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma_\beta^{-1} \Omega_\beta \Sigma_\beta^{-1}\right), \tag{3.3.4}$$

and, for all t*:*

$$\sqrt{N}(\widehat{\alpha}_{gt} - \alpha_{gt}^0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\omega_{gt}}{\pi_g^2}\right), \; g = 1, \dots, G^0, \tag{3.3.5}$$

where $\Sigma_\beta, \Omega_\beta, \omega_{gt}$*, and* π_g *are defined in Assumption 3.3.4.*

Consistent plug-in estimates of the asymptotic variances can easily be constructed (see, e.g., Supplemental Material in Bonhomme and Manresa, 2015).

3.3.3 Choice of the Preliminary Consistent Estimator

The interactive fixed effects literature provides a wide range of possible preliminary estimators for β. For instance, one may use (square root) nuclear norm penalized estimators (e.g., Beyhum and Gautier, 2019, 2023; Chernozhukov et al., 2019; Moon and Weidner, 2019) or robust methods Armstrong et al. (2022). Assuming knowledge of the number of group and a grouped factor structure on the covariates, one can alternatively use Chetverikov and Manresa (2021)'s spectral estimators.

To give a computationally very convenient example, we give conditions for using the nuclear norm estimator of Moon and Weidner (2019) as a preliminary consistent estimator of the slope coefficient.

Let $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_1$ denote the Frobenius norm and the nuclear norm respectively. Additionally, let $Y = (y_{it})_{i=1,\dots,N;t=1,\dots,T} \in \mathbb{R}^{N\times T}$, $X_k = (x_{it,k})_{i=1,\dots,N;t=1,\dots,T} \in$

$\mathbb{R}^{N\times T}$ for all $k \in \{1,\dots,p\}$, and $v \cdot X = \sum_{k=1}^{p} X_k v_k$ for all $v \in \mathbb{R}^p$. For all $(\psi, \beta')' \in (0,+\infty)\times\mathbb{R}^p$, define

$$Q_\psi(\beta) = \min_{\Gamma\in\mathbb{R}^{N\times T}} \left\{ \frac{1}{2NT} \|Y - \beta\cdot X - \Gamma\|_{\mathrm{F}}^2 + \frac{\psi}{\sqrt{NT}} \|\Gamma\|_1 \right\} \tag{3.3.6}$$

and

$$\widehat{\beta}^1(\psi) = \arg\min_{\beta\in\mathbb{R}^p} Q_\psi(\beta). \tag{3.3.7}$$

$\widehat{\beta}^1(\psi)$ is the nuclear norm regularized estimator (see, e.g., Moon and Weidner, 2019). It solves a very simple smooth and convex optimization problem (see, e.g. Hastie et al., 2015). Regularization is needed because the true number of groups is unknown. Define $\gamma^0 = (\mathbf{1}\{g_i^0 = g\})_{i=1,\dots,N;g=1,\dots,G^0} \in \{0,1\}^{N\times G^0}$, $\alpha^0 = (\alpha_{gt}^0)_{t=1,\dots,T;g=1,\dots,G^0} \in \mathcal{A}^{T\times G^0}$, $x_k = \mathrm{vec}(X_k)$, and $x = (x_1,\dots,x_k)$.

Assumption 3.3.5

(a) As N and T tend to infinity: $\psi \to 0$ such that $\sqrt{\min(N,T)}\psi \to \infty$.

(b) Let $\mathbb{C} = \left\{ A \in \mathbb{R}^{N\times T} : \left\| M_{\gamma^0} A M_{\alpha^0} \right\|_1 \le 3 \left\| A - M_{\gamma^0} A M_{\alpha^0} \right\|_1 \right\}$, where $M_B := I - B(B'B)^\dagger B'$, I is the identity matrix of appropriate dimensions, and † refers to the Moore-Penrose generalized inverse. There exists $\mu > 0$, independent from N and T, such that for any $a \in \mathbb{R}^{NT}$ with $\mathrm{mat}(a) \in \mathbb{C}$ we have $a' M_x a \ge \mu a' a$, for N,T sufficiently large.

(c) $\|(v_{it})_{i=1,\dots,N;t=1,\dots,T}\|_\infty = O_p\left(\sqrt{\max(N,T)}\right)$, where $\|\cdot\|_\infty$ denotes the spectral norm.

(d) As N and T tend to infinity: $\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T x_{it}x_{it}' \xrightarrow{p} \Sigma > 0$ and $\frac{1}{\sqrt{NT}}\sum_{i=1}^N\sum_{t=1}^T v_{it}x_{it} = O_p(1)$, where $\xrightarrow{p}$ denotes convergence in probability.

Assumption 3.3.5(b) is a restricted eigenvalue condition, common in high-dimensional modeling (see, e.g., Bickel et al., 2009). Sufficient conditions for Assumption 3.3.5(c) are given in the Supplementary Appendix S.2 of Moon and Weidner (2017).

Proposition 3.3.3 (Moon and Weidner (2019)) *Let Assumption 3.3.5 hold. Then, as N and T diverge jointly to infinity, $\left\|\widehat{\beta}^1(\psi) - \beta^0\right\| = O_p(\psi)$ with $\psi \to 0$.*

Proposition 3.3.3 follows from Theorem 2 in Moon and Weidner (2019), because of the interactive fixed effects structure of equation (3.3.1).

3.3.4 Choice of the Tuning Parameter

In finite samples, the parameter c_{NT} has to be chosen carefully. We propose a simple multiple fold cross-validation procedure which is asymptotically valid (not in the sense of minimizing the cross-validated ERM but in the sense of consistency and asympotic normality).

For simplicity suppose $\beta_0 = \widehat{\beta}^1 = 0$. Split the time dimension of the panel in K_1 distinct folds $\mathcal{T}_1, \ldots, \mathcal{T}_{K_1}$, and the cross-section dimension in K_2 distinct folds, $\mathcal{I}_1, \ldots, \mathcal{I}_{K_2}$. For $(k, \ell) \in \{1, \ldots, K_1\} \times \{1, \ldots, K_2\}$ and c_{NT} in a well-calibrated grid of values (such that any c_{NT} in this grid verifies Assumption 3.3.2),

- Step 1: estimate the group membership variables $\widehat{g}_1, \ldots, \widehat{g}_N$ using $(i, t) \in \{1, \ldots, N\} \times \mathcal{T}_\ell$.
- Step 2: estimate the group-specific time effects using $(i, t) \in \mathcal{I}_k \times \mathcal{T}_\ell^c$.
- Step 3: estimate the cross-validated MSE using $(i, t) \in \mathcal{I}_k^c \times \mathcal{T}_\ell^c$.

3.4 Discussion and Conclusion

Grouped fixed effects models are plagued with an underlying difficult combinatorial classification problem, rendering estimation and inference difficult. In this chapter, I propose a novel constructive identification argument for all the model parameters including the number of groups. The corresponding two-step estimator has polynomial computational cost and is straightforward to implement (only smooth convex optimization and elementary arithmetic operations are required). It is based on thresholding a suitable pairwise differencing transformation of the regression equation and a preliminary off-the-shelf consistent estimator of the slope. Mild conditions are given under which the proposed estimator is uniformly consistent for the latent grouping structure and asymptotically normal as both dimensions diverge jointly. Importantly, the number of groups is consistently estimated without any prior knowledge, and the time-dimension can grow much more slowly than the cross-sectional dimension. This work leaves a few questions unanswered. First, could the approach be fruitful to build a test for the grouping assumption? Second, how does the new estimator perform relative to alternative methods that require the number of groups to be known? Third, could similar differencing ideas be applied to more general nonlinear models? While the first two questions are left for further research, the last one is the object of Chapter 4.

3.5 Proofs of the Results

Define the matrix norm $\|\cdot\|_{\max}$ such that, for any $A = (a_{ij})_{i,j} \in \mathbb{R}^{m \times n}$, $\|A\|_{\max} = \max_{i=1,\ldots,m;j=1,\ldots,n} |a_{ij}|$.

3.5.1 Proof of Proposition 3.3.1

Let $W^0 = (\mathbf{1}\{g_i^0 = g_j^0\})_{i=1,\ldots,N;j=1,\ldots,N}$. Equations (3.3.2) and (3.3.3) are immediate corollaries of Lemma 3.5.1 below.

Lemma 3.5.1 *Let Assumptions 3.3.1-3.3.3 hold. Then, as N and T tend to infinity,*

$$\left\|\widehat{W} - W^0\right\|_{\max} = o_p(1). \tag{3.5.1}$$

Proof of Lemma 3.5.1: Let $\epsilon > 0$. By Assumption 3.3.1, there exists $K > 0$ such that, letting $\mathcal{E}_{1NT} = \left\{\left\|\widehat{\beta}^1 - \beta^0\right\| > Kr_{NT}\right\}$, $\mathbb{P}(\mathcal{E}_{1NT}) < \epsilon$ for N, T sufficiently large. Define $Z_{1NT}(i,j) = \widehat{W}_{ij}(1 - W^0_{ij})$, $Z_{2NT}(i,j) = (1 - \widehat{W}_{ij})W^0_{ij}$, and the probability events $\mathcal{E}_{2N} = \left\{\min_{g\in\{1,\dots,G^0\}} \sum_{i=1}^N \mathbf{1}\{g^0_i = g\} \geq 2\right\}$ and $\mathcal{E}_{NT} = \mathcal{E}^c_{1NT} \cap \mathcal{E}_{2N}$. By the union bound,

$$\begin{aligned}
&\mathbb{P}\left(\max_{(i,j)\in\{1,\dots,N\}^2} \left|\widehat{W}_{ij} - W^0_{ij}\right| > 0\right) \\
&\quad\leq \mathbb{P}(\mathcal{E}^c_{NT}) + \sum_{(i,j)\in\{1,\dots,N\}^2} \mathbb{P}\left(\left|\widehat{W}_{ij} \neq W^0_{ij}\right|, \mathcal{E}_{NT}\right) \\
&\quad\leq \mathbb{P}\left(\mathcal{E}_{1NT}\right) + \mathbb{P}\left(\mathcal{E}^c_{2NT}\right) + \sum_{(i,j)\in\{1,\dots,N\}^2} \mathbb{P}\left(\left|\widehat{W}_{ij} \neq W^0_{ij}\right|, \mathcal{E}_{NT}\right) \\
&\quad= \epsilon + o(1) + \sum_{(i,j)\in\{1,\dots,N\}^2} \mathbb{P}\left(Z_{1NT}(i,j) = 1, \mathcal{E}_{NT}\right) + \mathbb{P}\left(Z_{2NT}(i,j) = 1, \mathcal{E}_{NT}\right),
\end{aligned} \tag{3.5.2}$$

where I have used that $\lim_{N\to\infty} \mathbb{P}\left(\mathcal{E}^c_{2N}\right) = 0$ by Assumption 3.3.3(e) to obtain the equality. Below, I prove that, for $\ell \in \{1,2\}$, and as N and T tend to infinity,

$$\max_{(i,j)\in\{1,\dots,N\}^2} \mathbb{P}\left(Z_{\ell NT}(i,j) = 1, \mathcal{E}_{NT}\right) = o(N^2T^{-\delta}) \text{ for all } \delta > 0. \tag{3.5.3}$$

Equation (3.5.1) then follows by combining (3.5.2)-(3.5.3) and Assumption 3.3.2, and because ϵ is unrestricted.

1. I first show (3.5.3) for $\ell = 1$.[5] Let $(i,j) \in \{1,\dots,N\}^2$ and $\delta > 0$.

$$Z_{1NT}(i,j) = \mathbf{1}\Big\{\max_{k\in\{1,\dots,N\}\setminus\{i,j\}} \left|\frac{1}{T}\sum_{t=1}^T (\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt}\right| \leq c_{NT}\Big\}\mathbf{1}\{g^0_i \neq g^0_j\}.$$

If $G^0 = 1$, then almost surely $g^0_i = g^0_j$ and $Z_{1NT}(i,j) = 0$, i.e., (3.5.3) holds. Else,

$$\begin{aligned}
&\mathbf{1}\{Z_{1NT}(i,j) = 1, \mathcal{E}_{NT}\} \\
&\quad= \mathbf{1}\{\mathcal{E}_{NT}\}\times \\
&\qquad\sum_{\substack{(g,\widetilde{g})\in\{1,\dots,G^0\}^2 \\ g\neq\widetilde{g}}} \mathbf{1}\{g^0_i = g\}\mathbf{1}\{g^0_j = \widetilde{g}\}\mathbf{1}\Big\{\max_{k\in\{1,\dots,N\}\setminus\{i,j\}} \left|\frac{1}{T}\sum_{t=1}^T (\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt}\right| \leq c_{NT}\Big\}.
\end{aligned}$$

[5] Actually, I show the stronger result that the supremum is $o(T^{-\delta})$.

If $\mathbf{1}\{\mathcal{E}_{NT}\}\mathbf{1}\{g_i^0 \neq g_j^0\} = 1$, there exists a pair $\left(k^*(i,j,g_i^0), l^*(i,j,g_j^0)\right) \in \mathscr{P}_2\left(\{1,\dots,N\}\setminus\{i,j\}\right)$ such that $g^0_{k^*(i,j,g_i^0)} = g_i^0$ and $g^0_{l^*(i,j,g_j^0)} = g_j^0$. It follows that

$$
\begin{aligned}
&\mathbf{1}\{Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\}\\
&\quad\leq \mathbf{1}\{\mathcal{E}_{NT}\}\times\\
&\qquad\sum_{\substack{(g,\widetilde{g})\in\left\{1,\dots,G^0\right\}^2\\ g\neq\widetilde{g}}} \mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_j^0=\widetilde{g}\}\mathbf{1}\Big\{\left|\frac{1}{T}\sum_{t=1}^T(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{k^*(i,j,g_i^0)t}\right|\leq c_{NT}\Big\}\times\\
&\qquad\mathbf{1}\Big\{\left|\frac{1}{T}\sum_{t=1}^T(\widehat{v}_{it}-\widehat{v}_{jt})\widehat{v}_{l^*(i,j,g_j^0)t}\right|\leq c_{NT}\Big\}\\
&\quad\leq \mathbf{1}\{\mathcal{E}_{NT}\}\times\\
&\qquad\sum_{\substack{(g,\widetilde{g})\in\left\{1,\dots,G^0\right\}^2\\ g\neq\widetilde{g}}} \mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_j^0=\widetilde{g}\}\mathbf{1}\Big\{\left|\frac{1}{T}\sum_{t=1}^T(\widehat{v}_{it}-\widehat{v}_{jt})(\widehat{v}_{k^*(i,j,g_i^0)t}-\widehat{v}_{l^*(i,j,g_j^0)t})\right|\leq 2c_{NT}\Big\},
\end{aligned}
$$

where the first inequality uses the definition of the maximum, and the second inequality follows from the triangle inequality. Since there is at most one pair $(g,\widetilde{g})\in\{1,\dots,G^0\}^2$ such that $g\neq\widetilde{g}$ and $\mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_j^0=\widetilde{g}\}=1$, and by developing the product and using $\mathbf{1}\{|a|\leq b\}\leq\mathbf{1}\{a\leq b\}$ for any $(a,b)\in\mathbb{R}\times\mathbb{R}^*$, I obtain

$$
\begin{aligned}
&\mathbf{1}\{Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\}\\
&\quad\leq \mathbf{1}\{\mathcal{E}_{NT}\}\times\\
&\qquad\max_{\substack{(g,\widetilde{g})\in\left\{1,\dots,G^0\right\}^2\\ g\neq\widetilde{g}}} \mathbf{1}\Big\{\frac{1}{T}\sum_{t=1}^T\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2+\frac{1}{T}\sum_{t=1}^T\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(v_{it}-v_{jt}+v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^T\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(\beta^0-\widehat{\beta}^1\right)'\left(x_{it}-x_{jt}+x_{k^*(i,j,g)t}-x_{l^*(i,j,\widetilde{g})t}\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^T(v_{it}-v_{jt})\left(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^T\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt})\left(\beta^0-\widehat{\beta}^1\right)'\left(x_{k^*(i,j,g)t}-x_{l^*(i,j,\widetilde{g})t}\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^T(v_{it}-v_{jt})\left(\beta^0-\widehat{\beta}^1\right)'\left(x_{k^*(i,j,g)t}-x_{l^*(i,j,\widetilde{g})t}\right)\\
&\qquad+\frac{1}{T}\sum_{t=1}^T(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t})\left(\beta^0-\widehat{\beta}^1\right)'(x_{it}-x_{jt})\leq 2c_{NT}\Big\}\\
&\quad=\mathbf{1}\{\mathcal{E}_{NT}\}\times\max_{\substack{(g,\widetilde{g})\in\left\{1,\dots,G^0\right\}^2\\ g\neq\widetilde{g}}}\mathbf{1}\Big\{A_T(i,j,g,\widetilde{g})\leq 2c_{NT}\Big\},
\end{aligned}
\tag{3.5.4}
$$

where $A_T(i,j,g,\widetilde{g})$ is defined implicitely. I now define

$$\begin{aligned}B_T(i,j,g,\widetilde{g}) = \Bigg| & A_T(i,j,g,\widetilde{g}) - \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2 \\ & -\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)\left(v_{it} - v_{jt} + v_{k^*(i,j,g)t} - v_{l^*(i,j,\widetilde{g})t}\right) \\ & -\frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})\left(v_{k^*(i,j,g)t} - v_{l^*(i,j,\widetilde{g})t}\right)\Bigg|.\end{aligned}$$

Let $\overline{a} = \sup_{a\in\mathcal{A}}|a|$. By assumption 3.3.3(a), $\overline{a} < \infty$. It is easy to show using the Cauchy-Schwarz inequality that

$$\begin{aligned}B_T(i,j,g,\widetilde{g}) \leq \left\|\widehat{\beta}^1 - \beta^0\right\| & \left\{\frac{2\overline{a}}{T}\sum_{t=1}^{T}\left(\|x_{it}\| + \|x_{jt}\| + \left\|x_{k^*(i,j,g)t}\right\| + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|\right)\right. \\ & + \frac{4\left\|\widehat{\beta}^1 - \beta^0\right\|}{T}\sum_{t=1}^{T}\left(\|x_{it}\|^2 + \|x_{jt}\|^2 + \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2\right) \\ & + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{it}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2}\right)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2} \\ & \left. + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{k^*(i,j,g)t}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{l^*(i,j,\widetilde{g})t}^2}\right)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 + \|x_{jt}\|^2}\right\}.\end{aligned}$$

By Assumption 3.3.3(b), there exists $M^* > 0$ such that $\mathbb{E}(v_{it}^2) \leq M^*$ for all i,t. Let $\widetilde{M} > \max(M, \max(M^*, 1))$, where M is defined in Assumption 3.3.3(f) and let $\eta > 0$ such that

$$\eta \leq \min\left(1, \frac{c}{24\left(2\overline{a}4\sqrt{\widetilde{M}} + 8\widetilde{M} + 4\sqrt{2}\widetilde{M}\right)}\right). \tag{3.5.5}$$

Since $r_{NT} \to 0$ as $N,T \to \infty$, for N,T sufficiently large, $\left\|\widehat{\beta}^1 - \beta^0\right\| \leq \eta$ on $\mathcal{E}_{NT}$. Using the Cauchy-Schwarz inequality and $\eta \leq 1$, I obtain

$$\begin{aligned}\mathbf{1}\{\mathcal{E}_{NT}\}B_T(i,j,g,\widetilde{g}) \leq \eta & \left\{2\overline{a}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 + \|x_{jt}\|^2 + \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2}\right. \\ & + \frac{4}{T}\sum_{t=1}^{T}\left(\|x_{it}\|^2 + \|x_{jt}\|^2 + \left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2\right) \\ & + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{it}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{jt}^2}\right)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left\|x_{k^*(i,j,g)t}\right\|^2 + \left\|x_{l^*(i,j,\widetilde{g})t}\right\|^2} \\ & \left. + \left(\sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{k^*(i,j,g)t}^2} + \sqrt{\frac{1}{T}\sum_{t=1}^{T}v_{l^*(i,j,\widetilde{g})t}^2}\right)\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 + \|x_{jt}\|^2}\right\} \\ =: & \, C_T(i,j,g,\widetilde{g}).\end{aligned}$$

Plugging this bound into (3.5.4), I obtain

$$\begin{aligned}
&\mathbf{1}\{Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\}\\
&\leq \max_{\substack{(g,\widetilde{g})\in\{1,\dots,G^0\}^2\\ g\neq\widetilde{g}}} \mathbf{1}\Big\{\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\\
&+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)\left(v_{it}-v_{jt}+v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right)\\
&+\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})\left(v_{k^*(i,j,g)t}-v_{l^*(i,j,\widetilde{g})t}\right)\leq 2c_{NT}+C_T(i,j,g,\widetilde{g})\Big\}.
\end{aligned}$$

By the Cauchy-Schwarz inequality again, and because $\widetilde{M}\geq 1$, note the implication

$$\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\leq\widetilde{M}\implies\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\leq\sqrt{\widetilde{M}}\leq\widetilde{M}.$$

Using this result, the union bound, and some probability algebra, it follows that

$$\begin{aligned}
&\mathbb{P}\left(Z_{1NT}(i,j)=1,\mathcal{E}_{NT}\right)\\
&\leq\sum_{\substack{(g,\widetilde{g})\in\{1,\dots,G^0\}^2\\ g\neq\widetilde{g}}}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)v_{it}\leq-\frac{c}{12}+2c_{NT}+\eta\left(2\overline{a}4\sqrt{\widetilde{M}}+8\widetilde{M}+4\sqrt{2}\widetilde{M}\right)\right)\\
&+4G^0\left(G^0-1\right)\left[\sup_{g\neq\widetilde{g}}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\leq\frac{c}{2}\right)+\sup_{i\in\{1,\dots,N\}}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2\geq\widetilde{M}\right)\right.\\
&+\sup_{i\in\{1,\dots,N\},g\neq\widetilde{g}}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)v_{it}\right|\geq\frac{c}{12}\right)+\sup_{i\in\{1,\dots,N\}}\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\geq\widetilde{M}\right)\\
&\left.+\sup_{(i,j,k)\in\mathscr{P}_3(\{1,\dots,N\})}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it}-v_{jt})v_{kt}\right|\geq\frac{c}{12}\right)\right].
\end{aligned}\tag{3.5.6}$$

First, I bound the terms with a supremum. By Assumption 3.3.3(c), it holds that $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0)^2]=c_{g,\widetilde{g}}>c$. So for T large enough, I have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\right]\geq\frac{2c}{3}.$$

Applying Lemma B.5 in Bonhomme and Manresa (2015) to $z_t=(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0)^2-\mathbb{E}[(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0)^2]$, which satisfies appropriate mixing and tail conditions by Assumption 3.3.3(b) and (d), and taking $z=c/6$ yields, as T tends to infinity,

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\leq\frac{c}{2}\right)=o(T^{-\delta}),\tag{3.5.7}$$

uniformly across g and $\widetilde{g}$. Similarly, applying Lemma B.5 to $z_t = v_{it}^2 - \mathbb{E}(v_{it})^2$ and taking $z = \widetilde{M} - M^*$ yields

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T} v_{it}^2 \geq \widetilde{M}\right) = o(T^{-\delta}), \tag{3.5.8}$$

uniformly across units i. Note that $\{v_{it}^2\}_t$ is strongly mixing as $\{v_{it}\}$ is strongly mixing by Assumption 3.3.3(d). By Assumption 3.3.3(d), the process $\left\{(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0)v_{it}\right\}_t$ has zero mean, and is strongly mixing with faster-than-polynomial decay rate. Moreover, for all i, t and $m > 0$,

$$\mathbb{P}\left(\left|\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it}\right| > m\right) \leq \mathbb{P}\left(|v_{it}| > \frac{m}{2\overline{a}}\right),$$

so $\left\{(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0)v_{it}\right\}_t$ also satisfies the tail condition of Assumption 3.3.3(b), albeit with a different constant $b' > 0$ instead of $b > 0$. Lastly, applying Lemma B.5 from Bonhomme and Manresa (2015) again with $z_t = (\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0)v_{it}$ and taking $z = c/12$ yields

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it}\right| \geq \frac{c}{12}\right) = o(T^{-\delta}) \tag{3.5.9}$$

uniformly across i, g, and $\widetilde{g}$. An analogous reasoning yields

$$\sup_{(i,j,k)\in\mathscr{P}_3(\{1,\ldots,N\})} \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})v_{kt}\right| \geq \frac{c}{12}\right) = o(T^{-\delta}). \tag{3.5.10}$$

Finally, for N, T sufficiently large, $c_{NT} \leq c/72$ and a similar reasoning yields

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it} \leq -\frac{c}{12} + 2c_{NT} + \eta\left(2\overline{a}4\sqrt{\widetilde{M}} + 8\widetilde{M} + 4\sqrt{2}\widetilde{M}\right)\right) \\
&\leq \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right) v_{it} \leq -\frac{c}{72}\right) \\
&= o(T^{-\delta}),
\end{aligned} \tag{3.5.11}$$

uniformly across $g, \widetilde{g}$, where I have used the value of η given in (3.5.5). Combining (3.5.6)-(3.5.11) and using Assumption 3.3.3(f) yields

$$\sup_{(i,j)\in\{1,\ldots,N\}^2} \mathbb{P}\left(Z_{1NT}(i,j) = 1, \mathcal{E}_{NT}\right) = G^0(1 - G^0) \times o_p(T^{-\delta}) = o_p(T^{-\delta}),$$

i.e., (3.5.3) for $\ell = 1$ holds.

2. Second, I show (3.5.3) for $\ell = 2$. I now have

$$
\begin{aligned}
&\mathbf{1}\{Z_{2NT}(i,j) = 1, \mathcal{E}_{NT}\} \\
&\quad = \mathbf{1}\{\mathcal{E}_{NT}\}\mathbf{1}\Big\{ \max_{k\in\{1,\ldots,N\}\setminus\{i,j\}} \left| \frac{1}{T}\sum_{t=1}^{T}(\widehat{v}_{it} - \widehat{v}_{jt})\widehat{v}_{kt} \right| > c_{NT} \Big\} \mathbf{1}\{g_i^0 = g_j^0\} \\
&\quad \leq \mathbf{1}\{\mathcal{E}_{NT}\}\mathbf{1}\Big\{ \max_{k\in\{1,\ldots,N\}\setminus\{i,j\}} \left| \frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})v_{kt} + \frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})\alpha_{kt}^0 \right. \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T}\left(\beta^0 - \widehat{\beta}^1\right)'(x_{it} - x_{jt})\left(\beta^0 - \widehat{\beta}^1\right)' x_{kt} \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})\left(\beta^0 - \widehat{\beta}^1\right)' x_{kt} + \frac{1}{T}\sum_{t=1}^{T}\alpha_{kt}^0\left(\beta^0 - \widehat{\beta}^1\right)'(x_{it} - x_{jt}) \\
&\qquad \left. + \frac{1}{T}\sum_{t=1}^{T} v_{kt}\left(\beta^0 - \widehat{\beta}^1\right)'(x_{it} - x_{jt}) \right| > c_{NT} \Big\}.
\end{aligned}
$$

By the union bound, the triangle inequality, and the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&\mathbb{P}\left(Z_{2NT}(i,j) = 1, \mathcal{E}_{NT}\right) \\
&\quad \leq (N-2) \sup_{(i,j,k)\in\{1,\ldots,N\}^3} \left\{ \mathbb{P}\left(\left| \frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})v_{kt} \right| > \frac{c_{NT}}{10} \right) \right. \\
&\qquad + \mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2 + \|x_{jt}\|^2 + \|x_{kt}\|^2 > \frac{c_{NT}}{10 \times 4K^2 r_{NT}^2} \right) \\
&\qquad + 4\mathbb{P}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{it}^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|x_{kt}\|^2} > \frac{c_{NT}}{10 K r_{NT}} \right) \\
&\qquad + 2\mathbb{P}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{it}^2} > \frac{c_{NT}}{10 K r_{NT} \times \overline{a}} \right) \\
&\qquad \left. + 2\mathbb{P}\left(\sqrt{\frac{1}{T}\sum_{t=1}^{T} v_{it}^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\|x_{kt}\|^2} > \frac{c_{NT}}{10 K r_{NT} \times \overline{a}} \right) \right\}.
\end{aligned}
$$

Under the strong mixing and tail conditions given by Assumptions 3.3.3(b) and 3.3.3(d), and because $c_{NT}/r_{NT}^2 \to 0$ and $c_{NT}/r_{NT} \to 0$ by Assumption 3.3.2, all noninitial probabilities in the above expression can be shown to be $o(T^{-\delta})$ for all $\delta > 0$, uniformly accross (i,j,k), by similar arguments as in Step 1. For the first probability, a close inspection of the proof of Lemma B.5 in Bonhomme and Manresa (2015) reveals that, by taking $z_t = (v_{it} - v_{jt})v_{kt}$ and $z = c_{NT}/6$, and because $c_{NT} \gtrsim T^{-\kappa}$ as $N, T \to \infty$ by Assumption 3.3.2, for N, T sufficiently large,

$$
\begin{aligned}
\mathbb{P}\left(\left| \frac{1}{T}\sum_{t=1}^{T}(v_{it} - v_{jt})v_{kt} \right| \geq \frac{c_{NT}}{10} \right) \leq 4\left(1 + \frac{T^{1/2-2\kappa}}{C_1}\right)^{-(1/2)T^{1/2}} \\
+ C_2 T^{\kappa} \exp\left(-C_3\left(T^{(1/2-\kappa)/C_4}\right)\right), \quad (3.5.12)
\end{aligned}
$$

where C_1, C_2, C_3, and C_4 are positive constants that do not depend on i, j, k. Since $\kappa < 1/2$, the upper bound is $o_p(T^{-\delta})$ for all $\delta > 0$. This shows (3.5.3) for $\ell = 2$. □

The proof of Proposition 3.3.1 is complete.

3.5.2 Proof of Corollary 3.3.2

Let $\widetilde{\beta}$ and $(\widetilde{\alpha}_{11}, \ldots, \widetilde{\alpha}_{G^0T})'$ denote the infeasible oracle estimators computed using the pooled OLS regression of y_{it} on x_{it} and the interactions of group and time indicators $\mathbf{1}\{g_i^0 = 1\}, \ldots, \mathbf{1}\{g_i^0 = G^0\}$, $\mathbf{1}\{t = 1\}, \ldots, \mathbf{1}\{t = T\}$. By the same reasoning as in section S.A.1. in Bonhomme and Manresa (2015)'s Supplemental Material, I have

$$\sqrt{NT}(\widetilde{\beta} - \beta^0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \Sigma_\beta^{-1}\Omega_\beta\Sigma_\beta^{-1}\right), \tag{3.5.13}$$

and, for all $(g, t) \in \{1, \ldots, G^0\} \times \{1, \ldots, T\}$,

$$\sqrt{N}(\widetilde{\alpha}_{gt} - \alpha_{gt}^0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\omega_{gt}}{\pi_g^2}\right). \tag{3.5.14}$$

Without loss of generality, I assume that the chosen labels match the true group labeling. By Proposition 3.3.1, for all $(g, t) \in \{1, \ldots, G^0\} \times \{1, \ldots, T\}$,

$$\begin{aligned}
\mathbb{P}\left(\{\widehat{\alpha}_{gt} \neq \widetilde{\alpha}_{gt}\} \cup \left\{\widehat{\beta} \neq \widetilde{\beta}\right\}\right) &\leq \mathbb{P}\left(\widehat{G} \neq G^0\right) + \mathbb{P}\left(\max_{i \in \{1,\ldots,N\}} \left|\widehat{g}_i - g_i^0\right| > 0\right) \\
&= o(1) + o(1) \\
&= o(1).
\end{aligned}$$

Then, Eq. (3.3.5) follows from

$$\begin{aligned}
&\left|\mathbb{P}\left(\sqrt{N}(\widehat{\alpha}_{gt} - \alpha_{gt}^0) \leq a\right) - \mathbb{P}\left(\sqrt{N}(\widetilde{\alpha}_{gt} - \alpha_{gt}^0) \leq a\right)\right| \\
&\quad \leq \left|\mathbb{P}\left(\sqrt{N}(\widehat{\alpha}_{gt} - \alpha_{gt}^0) \leq a, \sqrt{N}(\widetilde{\alpha}_{gt} - \alpha_{gt}^0) > a\right)\right| \\
&\qquad + \left|\mathbb{P}\left(\sqrt{N}(\widehat{\alpha}_{gt} - \alpha_{gt}^0) > a, \sqrt{N}(\widetilde{\alpha}_{gt} - \alpha_{gt}^0) \leq a\right)\right| \\
&\quad \leq \quad \mathbb{P}\left(\widehat{\alpha}_{gt} \neq \widetilde{\alpha}_{gt}\right) + \mathbb{P}\left(\widehat{\alpha}_{gt} \neq \widetilde{\alpha}_{gt}\right) = o(1).
\end{aligned}$$

for any $a > 0$. Eq. (3.3.4) follows from a similar argument.

Chapter 4

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

Mais, comme on le sait, ce qui frappe l'esprit capricieux du poète n'est pas toujours ce qui impressionne la masse des lecteurs. Or, tout en admirant, comme les autres admireront sans doute, les détails que nous avons signalés, la chose qui nous préoccupa le plus est une chose à laquelle bien certainement personne avant nous n'avait fait la moindre intention.

Alexandre Dumas, *Les Trois Mousquetaires, Préface*

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.

4.1 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; Bonhomme et al., 2022; Hsiao, 2014; Moon and Weidner, 2019; Wooldridge, 2010).[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]

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

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

[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, Arellano and Hahn

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 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 Bonhomme et al., 2022; Chen et al., 2021; Zeleneev, 2020). 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 chapter, 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 $i \in \{1, \dots, N\}$'s outcome $Y_{it} \in \mathcal{Y}$ at time period $t \in \{1, \dots, T\}$ given i's covariates history $X_{i1}, \dots, X_{it}$, cluster membership $g_i \in \{1, \dots, G\}$, and cluster-specific effect $\alpha_{g_i t}$ is such that

$$\mathbb{P}\left(Y_{it} = y | X_{i1}, \dots, X_{it}, g_i, \alpha_{g_i t}\right) = h\left(y, X_{it}'\beta + \alpha_{g_i t}\right), \tag{4.1.1}$$

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 4.2). 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 (4.1.1) 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 et al. (2022) (see their Assumption 2) to build test functions which identify the sequence of latent clusterings $\{g_1, \dots, g_N\}_{N \geq 1}$ 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",

(2007); Chen et al. (2021,?); Dhaene and Jochmans (2015); Fernández-Val and Weidner (2016); Hahn and Newey (2004).

[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).

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 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. Computation, however, can be cumbersome in large samples. In Chapter 3, I showed 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 (4.1.1), 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 3.3.1 in Chapter 3). Here, I instead propose a simple heuristic, namely Lloyd (1982)'s algorithm described in Section 4.4.3, and show that it performs well in various Monte Carlo experiments with moderate sample sizes and number of clusters (see Section 4.6). 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 et al. (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

[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.

[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.

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 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 et al. (2022)'s 2-step GFE, Fernández-Val and Weidner (2016)'s nonlinear TWFE, or Chamberlain (1980); Rasch (1960)'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 et al. (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 ($t = 1973, \ldots, 1994$). 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

[9] See, e.g., Hahn and Newey (2004), Arellano and Hahn (2007), and Chen et al. (2021).

[10] Note that only Bonhomme et al. (2022)'s estimator assumes a correctly specified model. This paper does not provide inference tools. Comparison with Chen et al. (2021)'s estimator is left for further research.

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 4.10.3).

Overall, the theoretical results broaden the scope of application of GFE estimators and clustering 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 chapter 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 et al., 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 et al. (2021) and Mugnier and Wang (2022), which does not necessarily require large support (see also Vytlacil and Yildiz, 2007). This chapter 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

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

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

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., Arkhangelsky and Imbens, 2018; Bester and Hansen, 2016). 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 to be a difficult and much less investigated problem that I address in this chapter. The closest papers to ours are Chen et al. (2021), Bonhomme et al. (2022), and Ando and Bai (2022). Chen et al. (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 et al. (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 work, 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). 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 chapter 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

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

[14] See, Boneva et al. (2015); Gao et al. (2020); Liu et al. (2020); Su et al. (2016, 2019); Wang and Su (2021); Zhang et al. (2019).

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; Lloyd, 1982; MacQueen, 1967; McLachlan and Peel, 2000).

In Section 4.2, I introduce the class of NGFE models. The main identification result is presented in Section 4.3. In Section 4.4, I propose a general M-estimation framework, develop semiparametric NGFE estimators, and discuss computational aspects. Section 4.5 provides large sample properties in semiparametric binary choice models. Section 4.6 presents Monte Carlo results. Section 4.7 contains the empirical application. Section 4.8 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 $\text{Supp}(U)$ the support of any random variable U.

4.2 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 $i \in \mathcal{N} := \{1, \ldots, N\}$ at time $t \in \mathcal{T} := \{1, \ldots, T\}$ chooses $Y_{it} \in \mathcal{Y}$ given her

[15] See, among others, Athey et al. (2021); Kock (2016); Kock and Tang (2019); Moon and Weidner (2019); Zeleneev (2020).

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

[17] Unbalanced panels can be accomodated easily under exogeneous attrition (i.e., missing-at-random). Endogeneous attrition is beyond the scope of this chapter. 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 4.10.3.

[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).

weakly exogeneous covariates $X_i^t := (X_{i1}', \ldots, X_{it}')'$, her unobserved cluster membership variable $g_i^0 \in \mathcal{G}^0 := \{1, \ldots, G^0\}$, and unobserved time-varying cluster-specific effect $\alpha_{gt}^0 \in \mathcal{A} \subset \mathbb{R}$ such that, for all $y \in \mathcal{Y}$,

$$\mathbb{P}\left(Y_{it} = y | X_{i1}, \ldots, X_{it}, g_i^0, \alpha_{g_i^0 t}^0\right) = h^0\left(y, X_{it}'\beta^0 + \alpha_{g_i^0 t}^0\right), \tag{4.2.1}$$

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 $(\alpha_{g^0 t}^0)_{t \geq 1}$ 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 $\alpha_{g_i^0 t}^0$ enter the response function as the combination of a linear single-index $X_{it}'\beta^0 + \alpha_{g_i^0 t}^0$ and an unknown 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 (4.2.1), 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 7 (Binary outcome)

$$Y_{it} = \mathbb{1}\{X_{it}'\beta^0 + \alpha_{g_i^0 t}^0 - \varepsilon_{it} \geq 0\},$$

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

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

[19] If h^0 were known to the econometrician, model (4.2.1) would become a special case of the semiparametric nonlinear factor models considered in Chen et al. (2021).

Example 8 (Ordered outcome)

$$Y_{it} = \begin{cases} 0 & \text{if } X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} < d_1^0, \\ 1 & \text{if } d_1^0 \leq X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} < d_2^0, \\ 2 & \text{if } X_{it}'\beta^0 + \alpha^0_{g_i^0 t} - \varepsilon_{it} \geq d_2^0, \end{cases} \tag{4.2.2}$$

where $d_2^0 > d_1^0$*, and* ε_{it} *is independent from* $\left(X_{i1}', \ldots, X_{it}', g_i^0, \alpha^0_{g_i^0 t}\right)'$ *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 9 (Count outcome) $\mathcal{Y} = \{0, 1, 2, \ldots\}$*. A Poisson parametrization specifies*

$$h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) = \frac{\left(\lambda_{it}^0\right)^y \exp\left(-\lambda_{it}^0\right)}{y!}, \tag{4.2.3}$$

where $\lambda_{it}^0 = \exp\left(X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right)$*. 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, $\alpha^0 := (\alpha_{11}^0, \ldots, \alpha_{1T}^0, \ldots, \alpha_{G^0 1}^0, \ldots, \alpha_{G^0 T}^0)' \in \mathcal{A}^{G^0 T}$, and latent clustering structure $\gamma^0 := (g_1^0, \ldots, g_N^0)' \in \mathcal{G}^{0N}$. 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

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

While $\mathcal{B}$ is a finite-dimensional space, $\mathcal{H}$ is clearly not and the dimensions of both the discrete set $\{1, \ldots, G\}^N$ and $\mathcal{A}^{GT}$ grow with the sample size. This makes model (4.2.1) a high-dimensional combinatorial semi-parametric nonseparable model.

Remark 4.2.1 *It is straightforward to adapt the analysis to allow for cluster-specific slope coefficient* $\boldsymbol{\beta}^0 := \left(\beta_1^{0\prime}, \ldots, \beta_{G^0}^{0\prime}\right)'$ *such that*

$$\mathbb{P}\left(Y_{it} = y | X_{i1}, \ldots, X_{it}, g_i^0, \alpha^0_{g_i^0 t}, \beta^0_{g_i^0}\right) = h^0\left(y, X_{it}'\beta^0_{g_i^0} + \alpha^0_{g_i^0 t}\right), \quad \forall y \in \mathcal{Y}. \tag{4.2.4}$$

I discuss this extension, as well as heterogeneous link functions, additional individual- and time-specific effects, and grouped time-periods in Appendices 4.10.1-4.10.3. Model

(4.2.1) *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 4.2.2 *Model* (4.2.1) *extends Bonhomme and Manresa (2015) to nonparametric discrete choice modeling. In contrast to Bonhomme et al. (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.*

4.3 Large-N, Large-T Nonparametric Identification

In this section, I prove the nonparametric identification of θ^0_{NT} in model (4.2.1) 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 (4.2.1) is related to nonseparable panel data models with latent factors as it implies the following semiparametric regression equations:

$$\mathbb{1}\{Y_{it} = y\} = h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) + \varepsilon_{it}(y), \quad \forall (i,t,y) \in \mathcal{N} \times \mathcal{T} \times \mathcal{Y}, \tag{4.3.1}$$

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

$$Y_{it} = \sum_{y \in \mathcal{Y}} y h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 t}\right) + v_{it}, \quad \forall (i,t) \in \mathcal{N} \times \mathcal{T}, \tag{4.3.2}$$

where $v_{it} = \sum_{y \in \mathcal{Y}} y\varepsilon_{it}(y)$ and, by linearity, $\mathbb{E}\left[v_{it}|X_i, g_i^0, \alpha^0_{g_i^0 t}\right] = 0$. The representation given by (4.3.1) is useful to identify the clustering structure, while the representation given by (4.3.2) 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 $\|\beta^0\| = 1$ and $\alpha^0_{11} = 0$, where $\|\cdot\|$ denotes the Euclidean norm.[21] Identification is based on Assumptions 4.3.1-4.3.5 below.

Assumption 4.3.1 (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 $m \in (0,1)$.

[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. Botosaru and Muris, 2017b; Ichimura, 1993).

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

Assumptions 4.3.1(a)-4.3.1(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 4.3.1(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 4.10.1, I discuss several extensions such as cluster-specific slopes, individual-fixed and time-fixed effects which possibly affect all observed variables.[22]

Assumption 4.3.2 (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* $\rho_i(x) : \mathcal{X}^0 \ni x \mapsto \mathbb{P}\left(Y_{i2} = y | X_{i2} = x, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right)$, $\phi\left(\rho_i, \rho_j\right) = \mathbb{1}\{g_i^0 = g_j^0\}$.

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

Assumption 4.3.2(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 et al. (2022) which ensures that latent variables are recoverable from observed moments and leaves flexibility to the researcher for defining clusters of unobserved heterogeneity. In Appendix 4.9.2, I provide sufficient conditions for Assumption 4.3.2(a) to hold, including smoothness and the existence of a special regressor à la Honore 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 $\alpha^0_{g2}|X_{i2} = x, g_i^0, \mu_i^0, \xi_i^0$ and $\alpha^0_{g2}|X_{j2} = x, g_j^0, \mu_j^0, \xi_j^0$ 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, $\phi(f,g) = \mathbb{1}\{f = g\}$ 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 4.3.2(b) rules out asymptotically negligible clusters. Notice that allowing for an increasing number of clusters or negligible clusters would require substantial changes to Assumption 4.3.1 (e.g., as the cross-sectional identical distribution would not hold anymore). Note also that Assumption 4.3.2(a) could be generalized to be based instead on the (possibly infinite dimensional) full conditional distribution of the outcome.

Assumption 4.3.3 (Regularity and smoothness)

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

[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).

(b) Almost surely, $\mathbb{E}\left(\|X_{12}\|^2 |g_1^0, \alpha^0, \lambda^0, \mu^0\right)$ *is finite and* $\mathbb{E}\left(X_{12}X_{12}'|g_1^0, \alpha^0, \lambda^0, \mu^0\right)$ *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* $X_{it}'\beta^0 + \alpha^0_{g_i^0 t}$.

Assumption 4.3.3 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 4.3.4 (Monotonicity) *There exists* $y \in \mathcal{Y}$ *such that* $h^0(y,v)$ *is strictly monotonic in* v.

Assumption 4.3.4 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. Bonhomme et al., 2022; Chen et al., 2021) that impose strongly unimodal densities (see Ibragimov, 1956).

Assumption 4.3.5 (Compensating variations) *For all fixed* $(g, \widetilde{g}, t)$, *there exist* $x_1, x_2 \in \mathcal{X}$ *such that*

$$\alpha^0_{\widetilde{g}t} + x_1'\beta^0 = \alpha^0_{gt} + x_2'\beta^0. \tag{4.3.3}$$

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

$$\alpha^0_{g\widetilde{t}} + x_3'\beta^0 = \alpha^0_{gt} + x_4'\beta^0. \tag{4.3.4}$$

Assumption 4.3.5 requires sufficient variation in the covariates and has the same flavor as the *compensating variations* used in D'Haultfoeuille et al. (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 4.3.1 below is the main identification result. Let $W_N^0 = \left(\mathbb{1}\{g_i^0 = g_j^0\}\right)_{(i,j)\in\{1,\dots,N\}^2}$.

Theorem 4.3.1 *Let Assumptions 4.3.1-4.3.3(a) hold, and let* N *and* T *diverge jointly to infinity. Then,*

1. $(W_N^0)_{N\in\mathbb{N}^*}$ *and* G^0 *are point identified.*
2. *If Assumptions 4.3.3(b)-4.3.5 further hold, then* h^0, β^0, *and* $(\alpha^0_{gt})_{(g,t)\in\mathcal{G}^0\times\mathbb{N}^*}$ *are point identified.*

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

For the proof see Appendix 4.9.1.

Remark 4.3.1 *A key argument of the proof of Theorem 4.3.1 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_N^0 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 $g_i^0 = g_j^0$. Given the clustering structure of the model, note that W_N^0 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 $\mathcal{C}^0(i) := \left\{j \in \mathcal{N} : g_j^0 = g_i^0\right\}$ 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 $y \in \mathcal{Y}$.*

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 et al. (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.

4.4 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

[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).

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.

4.4.1 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

$$\theta^0_{NT} = \underset{\theta \in \mathcal{B}\times\mathcal{H}\times\mathcal{G}^{0N}\times\mathcal{A}^{G^0T}}{\arg\max} \mathbb{E}\left(\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho(Y_{it}, X_{it}; \theta) | \gamma, \alpha \right), \tag{4.4.1}$$

where $\mathcal{G}^{0N} = \{1, \dots, G^0\}^N$ is the set of all partitions of $\{1, \dots, N\}$ 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

$$\widehat{\theta}^{\mathrm{M}}_{\rho} \in \underset{\theta \in \mathcal{B}\times\mathcal{H}\times\mathcal{G}^{0N}\times\mathcal{A}^{G^0T}}{\arg\max} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho(Y_{it}, X_{it}; \theta). \tag{4.4.2}$$

Finding a suitable ρ-function, proving identification of θ^0_{NT} (i.e., that Eq. (4.4.1) holds), and deriving the asymptotic properties of the sequence of $\widehat{\theta}^{\mathrm{M}}_{\rho}$ are certainly difficult problems beyond the scope of the chapter, each of them would require further assumptions. Moreover, computation of $\widehat{\theta}^{\mathrm{M}}_{\rho}$ is generally infeasible because maximization problem (4.4.2) 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 chapter. Instead, I focus on semiparametric versions where h^0 is assumed to be known and that are of practical interest in many empirical applications.

4.4.2 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 (4.2.1) 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 $\widehat{\theta}^{\mathrm{M}}_{\rho}$ (once suppressing dependence on h) with $\rho(Y_{it}, X_{it}; \theta) = \ln h^0 (Y_{it}, X'_{it}\beta + \alpha_{g_i t})$. In other words, $\widehat{\theta}^{\mathrm{NGFE}}$ is solution to

[26] Estimating G^0 in nonlinear models with time-varying unobserved heterogeneity is a difficult problem that is beyond the scope of the chapter. See Chen et al. (2021) for a discussion in some concave nonlinear factor type models. An AIC or BIC-type criterion à la Bai and Ng (2002); Bonhomme and Manresa (2015) 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 chapter. Note that Bonhomme et al. (2022) *need* the number of clusters to increase as they assume a (possibly) continuous underlying unobserved heterogeneity.

the following minimization problem:

$$\widehat{\theta}^{\text{NGFE}} \in \underset{\theta\in\mathcal{B}\times\mathcal{G}^{0N}\times\mathcal{A}^{G^0T}}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln h^0\left(Y_{it}, X_{it}'\beta+\alpha_{g_it}\right), \tag{4.4.3}$$

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

$$\widehat{g}_i(\beta,\alpha) = \underset{g\in\mathcal{G}^0}{\arg\min} \frac{1}{NT}\sum_{t=1}^{N}\sum_{t=1}^{T} -\ln h^0\left(Y_{it}, X_{it}'\beta+\alpha_{gt}\right), \tag{4.4.4}$$

where the minimum g is taken in case of a non-unique solution. The NGFE estimator of $(\beta^{0\prime}, \alpha^{0\prime})'$ in (4.4.3) can then be written as

$$(\widehat{\beta},\widehat{\alpha}) = \underset{(\beta,\alpha)\in\mathcal{B}\times\mathcal{A}^{G^0T}}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln h^0\left(Y_{it}, X_{it}'\beta+\alpha_{\widehat{g}_i(\beta,\alpha)t}\right), \tag{4.4.5}$$

where $\widehat{g}_i(\beta,\alpha)$ is given by (4.4.4).

4.4.3 Computation

The minimization problem (4.4.3) 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 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 $(\beta^{(0)}, \alpha^{(0)}) \in \mathcal{B}\times\mathcal{A}^{G^0T}$ be some starting value. Set $s = 0$.

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

$$g_i^{(s+1)} = \underset{g\in\mathcal{G}^0}{\arg\min} \sum_{t=1}^{T} -\ln h^0\left(Y_{it}, X_{it}'\beta^{(s)}+\alpha_{gt}^{(s)}\right). \tag{4.4.6}$$

3. Compute:

$$\left(\beta^{(s+1)}, \alpha^{(s+1)}\right) = \underset{(\beta,\alpha)\in\mathcal{B}\times\mathcal{A}^{G^0T}}{\arg\min} \sum_{i=1}^{N}\sum_{t=1}^{T} -\ln h^0\left(Y_{it}, X_{it}'\beta+\alpha_{g_i^{(s+1)}t}\right). \tag{4.4.7}$$

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

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

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]

4.5 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 4.3.1) 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 4.10.4), 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. (4.4.3).

4.5.1 Binary Choice Model With Grouped Fixed Effects

Consider the following data generating process:

$$Y_{it} = \mathbb{1}\{X_{it}'\beta^0 + \alpha_{g_i^0 t}^0 - \varepsilon_{it} \geq 0\}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T. \tag{4.5.1}$$

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 4.5.1
Eq. (4.5.1) *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*.*

[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.

[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.

Assumption 4.5.1(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 4.5.1(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 Bonhomme et al., 2022; Chen et al., 2021; Fernández-Val and Weidner, 2016).[30] The second part of Assumption 4.5.1(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 4.5.1, note that Eq. (4.5.1) is a semiparametric NGFE model (4.2.1) with known link function $h^0(y,z) = \Psi(z)^{\mathbb{1}\{y=1\}}(1-\Psi(z))^{\mathbb{1}\{y=0\}}$. The corresponding NGFE estimator writes

$$(\widehat{\beta}, \widehat{\gamma}, \widehat{\alpha}) \in \underset{(\beta,\gamma,\alpha)\in\mathcal{B}\times\mathcal{G}^{0N}\times\mathcal{A}^{G^0T}}{\arg\min} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{g_it}\right)\right), \tag{4.5.2}$$

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

4.5.2 Consistency

Consider the following assumption.

Assumption 4.5.2

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

(b) There exists a constant $M > 0$ such that $\|X_{it}\| \leq M$ 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, $\operatorname{plim}_{N,T\to\infty}\min_{\gamma\in\Gamma_{G^0}}\widehat{\rho}(\gamma) = \rho > 0$.

Assumption 4.5.2(a) is standard in the context of M-estimation. Assumption 4.5.2(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 4.5.2(a)) strong concavity of the log-likelihood function and rules non-stationary covariates.[31] Assumption 4.5.2(c) is the same noncollinearity condition as Assumption 1(g) in

[30] See also, Pratt (1981).

[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.

Bonhomme and Manresa (2015). It requires that X_{it} shows sufficient within-cluster variation over time and across individuals, and is related to standard noncollinearity assumptions encountered in the large-N, large-T panel data literature (see, e.g., Ando and Bai, 2022; Bai, 2009; Chen et al., 2021; Vogt and Linton, 2017). 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 4.5.2(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 support.

Theorem 4.5.1 (Consistency) *Let Assumptions 4.5.1 and 4.5.2 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 4.9.3.

Theorem 4.5.1 shows that NGFE estimators of the common slope coefficient and cluster-specific effects in NGFE binary choice models are both consistent.

4.5.3 Asymptotic Distribution

Consider the following assumption.

Assumption 4.5.3

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

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

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

Assumptions 4.5.3(a)-(c) are identical to Assumptions 2(a)-(c) in Bonhomme and Manresa (2015), respectively. Assumption 4.5.3(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 et al., 2021). Assumption 4.5.3(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 4.5.3(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 4.5.3 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 4.5.1, 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_i^0:

$$(\widetilde{\beta}, \widetilde{\alpha}) = \underset{(\beta,\alpha)\in\mathcal{B}\times\mathcal{A}^{G^0T}}{\operatorname{argmin}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{g_i^0 t}\right)\right). \tag{4.5.3}$$

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 4.5.1, 4.5.2, and 4.5.3 provide conditions under which estimated cluster memberships converge to their population counterparts, and the NGFE estimator defined in (4.5.2) 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 4.9.7 in Appendix 4.9.4). 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 4.5.1 and Eq. (4.9.27) 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 4.9.7, showing asymptotic normality of the NGFE estimator then reduces to the 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 $(-\ln\Psi)''(Q_{it}Z_{it}^0)$:

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

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:

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

Assumption 4.5.4 below allows to characterize the asymptotic distribution of the infeasible MLE $(\widetilde{\beta}, \widetilde{\alpha})$.

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

Assumption 4.5.4

(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)*:*

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

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

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

(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 4.5.4(a) rules out dynamic or feedbacks.

Theorem 4.5.2 (Asymptotic distribution) *Let Assumptions 4.5.1-4.5.4 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:*

$$\sqrt{NT}(\widehat{\beta} - \beta^0) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta^{-1}\right), \tag{4.5.4}$$

and, for all (g,t)*,*

$$\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha_{gt}^0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\widetilde{\pi}_{gt}^2}\right), \tag{4.5.5}$$

where Σ_β*,* ω_{gt}*, and* $\widetilde{\pi}_{gt}$ *are defined in Assumption 4.5.4.*

For the proof see Appendix 4.9.4.

Theorem 4.5.2 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 $T/N \to 0$ 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. Ando and Bai,

2022; Bai, 2003, 2009) 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 Chen et al. (2021); Fernández-Val and Weidner (2016). The intuition behind this result is that the extreme sparsity of the factor loading structure in model (4.5.1) 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 4.11.

4.5.4 Average Partial Effects (APEs)

Under Assumption 4.5.1, 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 $\delta^0_{NT} = \frac{1}{T}\sum_{t=1}^T \delta^0_t$ 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).

4.6 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

[33] To see the factor-loading structure, note that model (4.5.1) can be written as $Y_{it} = \mathbb{1}\{X'_{it}\beta + \lambda'_i f_t - \varepsilon_{it} \geq 0\}$, where $\lambda'_i = (\mathbb{1}\{g_i^0 = 1\}, \ldots, \mathbb{1}\{g_i^0 = G^0\}) \in \left\{b \in \{0,1\}^{G^0} : \sum_{g=1}^{G^0} b_g = 1\right\}$ 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 et al. (2021) apply and bias-correction methods are needed.

et al. (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.

4.6.1 Static Logit Model

The data generating process is

$$Y_{it} = \mathbb{1}\{X_{it}\beta + \alpha_{g_i t} > \varepsilon_{it}\}, \quad i = 1, \dots, N,\ t = 1, \dots, T, \tag{4.6.1}$$

where $\beta = 1$ and $\varepsilon_{it} \sim$Logit$(0, \pi^2/3)$, $g_i \sim$Unif$\{1, \dots, G^0\}$ 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 $\mu = (-2\pi/\sqrt{3}, -\pi/\sqrt{3}, 0, \pi/\sqrt{3}, 2\pi/\sqrt{3})'$ if $G^0 = 5$, V_i such that $\mathbb{P}(V_i = -2) = 1/12, \mathbb{P}(V_i = -1) = 1/4, \mathbb{P}(V_i = 0) = 1/3, \mathbb{P}(V_i = 1) = 1/4, \mathbb{P}(V_i = 2) = 1/12$, and $W_{it} \sim \mathcal{N}(0, 1)$:

- DGP 1 (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$[0, 1]$, $X_{it} = 0.5V_i + 0.8U_{g_i^0 t}$.
- DGP 2 (grouped patterns of time-invariant UH): $\alpha_{gt} = \mu_g$, $X_{it} = 0.3\mu_{g_i} + V_i + 0.8W_{it}$.
- DGP 3 (continuous time-invariant UH): $\alpha_i \sim \mathcal{N}(0, 1)$, $X_{it} = \alpha_i + 0.5V_i + 0.8W_{it}$.
- DGP 4 (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).

[34] I leave comparison with Charbonneau (2017)'s conditional logit and Chen et al. (2021)'s nonlinear factor models for further research. A definition of the metrics and more details are given in Appendix 4.12.

Table 4.1 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 4.2 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 et al. (2022)'s 2-step GFE.

Table 4.3 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 4.3 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 $G^0 = 2$).

4.6.2 Dynamic Logit Model

The data generating process is

$$\begin{aligned} Y_{it} &= \mathbb{1}\{Y_{it-1}\beta_1 + X_{it}\beta_2 + \alpha_{g_it} > \varepsilon_{it}\}, \quad i = 1, \dots, N,\ t = 1, \dots, T, \\ Y_{i0} &= \mathbb{1}\{X_{i0}\beta_2 + \alpha_{g_i0} > \varepsilon_{i0}\}, \quad i = 1, ..., N, \end{aligned} \tag{4.6.2}$$

where $\beta_1 = 0.5$ and $\beta_2 = 1$. Tables 4.4-4.6 report the same statistics as Tables 4.1-4.3 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 4.5.4(a) does not hold here). Previous comparisons still apply there.

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

4.7 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., Gilbert, 2006; Griffith and Van Reenen, 2021).[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 et al. (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 $p \in \{0, 1, \ldots\}$

$$\begin{aligned}&\mathbb{P}(\text{cwpatent}_{it} = p|\text{comp}_{it}, \nu_i, \xi_t)\\ &\quad = \frac{\exp(p(g(\text{comp}_{it}) + \nu_i + \xi_t)\exp(-\exp(g(\text{comp}_{it}) + \nu_i + \xi_t)))}{p!},\end{aligned} \tag{4.7.1}$$

where cwpatent_{it} represents the number of citation-weighted patents in industry i in year t, comp_{it} 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 4.1 shows ABBGH's original inverted-U relationship, by replicating ABBGH's Figure II, a scatterplot comparing the fit of the exponential model (4.7.1) with that of a nonparametric spline.[38]

While model (4.7.1) is in line with a large body of the previous literature (see, e.g., Gourieroux et al., 1984; Hausman et al., 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

[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.

[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 4.7. See Aghion et al. (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).

fragile as recent empirical research has reported both increasing and decreasing monotonic relationships depending on the controls included (Aghion et al., 2013), whether accounting or not for the presence of structural breaks (Correa, 2012), or the country data used (Askenazy et al., 2013; Hashmi, 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 semi-parametric NGFE estimators introduced in this chapter 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 are missing-at-random.[41] Table 4.7 reports summary statistics borrowed from Hashmi (2013). In particular, one can see that some industries are never granted patents.[42] Table 4.8 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

[39] Fernández-Val and Weidner (2016) estimate model (4.7.1), 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.

[41] While the time dimension is large, the cross-sectional dimension is slightly at odd with the asymptotic framework I consider. 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 et al., 2005).

the pairwise differencing estimator developed in Chapter 3 to ABBGH's residuals:

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

plotted in Figure 4.2. 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 4.3 and Figure 4.4 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 4.4 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:

$$\begin{aligned}&\mathbb{P}(\text{cwpatent}_{it} = p|\text{comp}_{it}, g_i, \alpha_{g_it}) \\ &\quad = \frac{\exp(p(g(\text{comp}_{it}) + \alpha_{g_it})\exp(-\exp(g(\text{comp}_{it}) + \alpha_{g_it})))}{p!}, \quad \forall p \in \{0, 1, \ldots\},\end{aligned} \tag{4.7.2}$$

where $g_i \in \{1, \ldots, G\}$ is industry i's unknown cluster membership and $(\alpha_{1t}, \ldots, \alpha_{Gt})' \in \mathbb{R}^G$ 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 (4.7.1) and (4.7.2) are non-nested as $G << N$.

Table 4.9 and Figure 4.5 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 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.

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

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 4.6 confirms this intuition by plotting the estimated cluster-specific effects obtained in specifications (3)-(5) from Table 4.9, where the data-driven clustering of industries is displayed in Figure 4.7.

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 4.8 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

4.10 show that coefficient estimates are similar to Table 4.9 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 AB-BGH'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 4.10).

4.8 Conclusion

In this chapter, 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 et al. (2005), I demonstrate that the so-called inverted-U relationship between innovation and product market competition is sensitive to the researcher's choice of whether controlling 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 Chapter 3, I propose such an estimator for linear versions of NGFE models and obtain this result under relatively weak conditions (see Proposition 3.3.1). 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 chapter. I leave such extensions for future work.

4.9 Proofs 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 $V_n = o_p(1)$.

4.9.1 Proof of Theorem 4.3.1

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 4.3.2, 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 $W_{ijN}^0 := \mathbb{1}\{g_i^0 = g_j^0\} = \phi(\rho_i, \rho_j)$ with $\rho_i(x) : \mathcal{X}^0 \ni x \mapsto \mathbb{P}(Y_{i2} = \overline{y} | X_{i2} = x, g_i^0, \mu_i^0, \xi_i^0)$. 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 4.3.1(b) and 4.3.3(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 $\mathbb{E}\left[\mathbb{1}\{Y_{it} = \overline{y}\} | X_{i2} = x, g_i^0, \mu_{g_i^0}^0, \xi_i^0\right]$. Hence, point identification of $\mathbb{E}\left[\mathbb{1}\{Y_{i2} = \overline{y}\} | X_{i2} = x, g_i^0, \mu_{g_i^0}^0, \xi_i^0\right] = \rho_i(x)$ 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 4.3.1(a) and 4.3.2(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, $\mathcal{C}^0(i) := \left\{j \in \{1,\ldots,N\} : g_j^0 = g_i^0\right\}$ is identified for all $N \in \mathbb{N}^*$. Under Assumption 4.3.1(a) and 4.3.2(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 (1993) with $\varphi(\cdot) = \sum_{y\in\mathcal{Y}} y h^0(y, \cdot + \alpha_{g_i^0 t}^0)$, whose conditions 4.1 and 4.2(1-3) hold under Assumptions 4.3.1(c) and 4.3.3, β^0 is identified up-to-scale. Because $\|\beta^0\| = 1$, β^0 is identified.

Identification of cluster-specific time effects α_{gt}^0 for all $(g,t) \in \mathcal{G}^0 \times \mathbb{N}^*$, up to cluster relabeling. Given identification of W_N^0 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

[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).

rows in W_N^0. 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 4.3.4. 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 4.3.1(a) and 4.3.1(c), the following probabilities are identified:

$$\mathbb{P}\left(Y_{1t} = \underline{y} | X_{1t} = x, g_1^0 = g, \alpha_{gt}^0\right) = h^0\left(\underline{y}, x'\beta^0 + \alpha_{gt}^0\right),$$
$$\mathbb{P}\left(Y_{1t} = \underline{y} | X_{1t} = x, g_1^0 = \widetilde{g}, \alpha_{\widetilde{g}t}^0\right) = h^0\left(\underline{y}, x'\beta^0 + \alpha_{\widetilde{g}t}^0\right).$$

By Assumption 4.3.5 (Eq. (4.3.3)), I can find $x_1, x_2 \in \mathcal{X}$ such that

$$\mathbb{P}\left(Y_{1t} = \underline{y} | X_{1t} = x_2, g_1^0 = g, \alpha_{gt}^0\right) = \mathbb{P}\left(Y_{1t} = \underline{y} | X_{1t} = x_1, g_1^0 = \widetilde{g}, \alpha_{\widetilde{g}t}^0\right),$$

or, equivalently,

$$h^0\left(\underline{y}, x_2'\beta^0 + \alpha_{gt}^0\right) = h^0\left(\underline{y}, x_1'\beta^0 + \alpha_{\widetilde{g}t}^0\right). \tag{4.9.1}$$

By strict monotonicity of $h^0(\underline{y}, \cdot)$, I can invert (4.9.1) and identify $\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0 = (x_2 - x_1)'\beta^0$. As β^0 is already identified, it follows that $\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0$ is identified. Because the result holds for all $(g, \widetilde{g}, t)$, it holds for $g = t = 1$ (for which $\alpha_{gt}^0 = 0$ by the normalization assumption), thus $(\alpha_{g1}^0)_{g \in \mathcal{G}^0}$ is identified. A similar reasoning but now identifying $x_1, x_2 \in \mathcal{X}$ such that Eq. (4.3.4) holds in place of Eq. (4.3.3) yields identification of $\alpha_{g\widetilde{t}}^0 - \alpha_{gt}^0$ for all $(g, t, \widetilde{t})$, and, in turn, that of $(\alpha_{1t}^0)_{t \in \mathbb{N}^*}$. Identification of α_{gt}^0 for all (g, t) then follows because, for all (g, t) with $g \neq 1$ and $t \neq 1$, α_{gt}^0 can be decomposed as

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

where a and b are identified. Finally, $h^0(y, z)$ is identified as a function of $y \in \mathcal{Y}$ and index $z = X_{it}'\beta^0 + \alpha_{g_i^0 t}^0$.
The proof of Theorem 4.3.1 is complete.

4.9.2 Sufficient Condition for Assumption 4.3.2(a)

Consider the following assumption.

Assumption 4.9.1

(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 $\alpha_{g2}^0 | X_{i2} = x, g_i^0 = g, \mu_{g_i^0}^0, \xi_i^0$ admits a fully supported density $f_{\alpha_{g2}^0 | X_{i2} = x, g_i^0 = g, \mu_g^0, \xi_i^0}(\alpha)$ with respect to the Lebesgue measure such that

$$f_{\alpha_{g2}^0 | X_{i2} = x, g_i^0 = g, \mu_g^0, \xi_i^0}(\alpha) = f_{\alpha_{\widetilde{g}2}^0 | X_{j2} = x, g_j^0 = \widetilde{g}, \mu_{\widetilde{g}}^0, \xi_j^0}(\alpha), \quad \lambda(\alpha)\text{-a.e.}$$

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

(b) *There exists* $k \in \{1,\dots,p\}$ *such that* $\beta_k^0 \neq 0$ *and* $X_{i2,k} \perp \alpha^0_{g_i^0 2} | X_{i2,(-k)}, g_i^0, \mu^0_{g_i^0}, \xi_i^0$. *Moreover, almost surely,* $Supp\left(X_{i2,k} | X_{i2,(-k)}, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right)$ *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* $f_\zeta(z) = \frac{|\psi_y'(z)|}{\int |\psi_y'| d\lambda}$ *does not vanish and is infinitely often differentiable in* $\mathbb{R}\backslash A$ *for some set* A *such that* $\lambda(A) = 0$.

Assumption 4.9.1(b) requires the existence of a special regressor (as in Honore and Lewbel, 2002), but (possibly) without large support (it depends on the support of the unobserved heterogeneity). Assumption 4.9.1(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 4.9.1(b) to be the full real line, which is equivalent to having a special regressor with large support à la Honore and Lewbel (2002).

Lemma 4.9.1 *If Assumptions 4.3.1(c) and 4.9.1 hold, then Assumption 4.3.2(a) holds.*

Proof of Lemma 4.9.1 W.l.o.g. I assume that $k = 1$ and denote $x_{(-1)} = (x_j)_{j\in\{2,\dots,p\}}$. Let $x = (x_1, x_{(-1)}')' \in \mathcal{X}^1$, and $y \in \mathcal{Y}$ verifying Assumption 4.9.1(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 4.3.2.
Step 1: Let $(i,x) \in \mathcal{N} \times \mathcal{X}^1$ and $\rho_i(x) := \mathbb{P}\left(Y_{i2} = y | X_{i2} = x, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right)$. By the law of total expectations, Assumption 4.3.1(c), using equation (4.2.1), and Assumption 4.9.1(a), I obtain

$$\begin{aligned}
\rho_i(x) &= \mathbb{E}\left[\mathbb{P}\left(Y_{i2} = y | X_{i2} = x, g_i^0, \alpha^0, \lambda^0, \mu^0, \xi^0\right) | X_{i2} = x, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(Y_{i2} = y | X_{i2} = x, g_i^0, \alpha^0_{g_i^0 2}\right) | X_{i2} = x, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right] \\
&= \mathbb{E}\left[\psi_y\left(x'\beta^0 + \alpha^0_{g_i^0 2}\right) | X_{i2} = x, g_i^0, \mu^0_{g_i^0}, \xi_i^0\right] \\
&= \int_{\mathbb{R}} \psi_y\left(x'\beta^0 + \alpha\right) f_{\alpha^0_{g_i^0 2} | X_{i2}=x, g_i^0, \mu^0_{g_i^0}, \xi_i^0}(\alpha) d\lambda(\alpha).
\end{aligned} \tag{4.9.2}$$

By Assumption 4.9.1(b), there exists $\epsilon > 0$ and an open set $\mathcal{X}^0 = \{x + (v, 0')' : v \in (-\epsilon, \epsilon)\} \subset \mathcal{X}^1$ with $\mathbb{P}\left(X_{i2} \in \mathcal{X}^0\right) > 0$ such that, for all $w \in \mathcal{X}^0$, almost everywhere $f_{\alpha^0_{g_i^0 2} | X_{i2}=w, g_i^0, \mu^0_{g_i^0}, \xi_i^0}(\alpha) = f_{\alpha^0_{g_i^0 2} | X_{i2}=x, g_i^0, \mu^0_{g_i^0}, \xi_i^0}(\alpha)$. Since $\mathcal{X}^0 \subset \mathcal{X}^1$,

[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).

Eq. (4.9.2) yields, for all $w \in \mathcal{X}^0$,

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

By Assumption 4.9.1(c), $w \mapsto \rho_i(w)$ is differentiable on $\mathcal{X}^0$ and, for all $w \in \mathcal{X}^0$,

$$\begin{aligned}\frac{\partial \rho_i(z_1,\ldots,z_p)}{\partial z_1}\Big|_{z=w} &= \beta_1^0 \int_{\mathbb{R}} \psi_y'\left(w'\beta^0 + \alpha\right) f_{\alpha^0_{g_i^0 2}|X_{i2}=x,g_i^0,\mu^0_{g_i^0},\xi_i^0}(\alpha)d\lambda(\alpha) \\ &= \beta_1^0\left(1 - 2\mathbb{1}\{\psi_y'(0) < 0\}\right) \\ &\quad \times \int_{\mathbb{R}} \left|\psi_y'\left(w'\beta^0 + \alpha\right)\right| f_{\alpha^0_{g_i^0 2}|X_{i2}=x,g_i^0,\mu^0_{g_i^0},\xi_i^0}(\alpha)d\lambda(\alpha),\end{aligned} \tag{4.9.3}$$

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 $\phi(f,g) := \mathbb{1}\{\Delta(\partial_1 f, \partial_1 g) = 0\}$ verifies Assumption 4.3.2(a). I have to show that, for all $(i,j) \in \mathcal{N}^2$,

$$\frac{\partial \rho_i(z_1,\ldots,z_p)}{\partial z_1}\Big|_{z=w} = \frac{\partial \rho_j(z_1,\ldots,z_p)}{\partial z_1}\Big|_{z=w} \quad \forall w \in \mathcal{X}^0 \iff g_i^0 = g_j^0. \tag{4.9.4}$$

Let $(i,j) \in \mathcal{N}^2$.
$\underline{\Longleftarrow}$: Suppose that $g_j^0 = g_i^0$ and let $w \in \mathcal{X}^0$. By Assumption 4.9.1(c), I have

$$f_{\alpha^0_{g_i^0 2}|X_{i2}=x,g_i^0,\mu^0_{g_i^0},\xi_i^0}(\alpha) = f_{\alpha^0_{g_j^0 2}|X_{j2}=x,g_j^0,\mu^0_{g_j^0},\xi_j^0}(\alpha), \quad \lambda(\alpha) - \text{a.e.}.$$

Equation (4.9.3) then implies $\frac{\partial \rho_i(z_1,\ldots,z_p)}{\partial z_1}\Big|_{z=w} = \frac{\partial \rho_j(z_1,\ldots,z_p)}{\partial z_1}\Big|_{z=w}$.
$\underline{\Longrightarrow}$: Suppose that, for all $w \in \mathcal{X}^0$,

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

Dividing each side of this equation by $\int \left|\psi_y'\right| d\lambda > 0$, using (4.9.3) and the fact that

$$\left|\left(1 - 2\mathbb{1}\{\psi_y'(0) < 0\}\right)\beta_1^0\right| = \left|\beta_1^0\right| > 0,$$

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

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

I show below that this constraint is equivalent to $f_{\alpha^0_{g_j^0}} = f_{\alpha^0_{g_i^0}}$ a.e., which, by Assumption 4.9.1(a), in turn implies $g_i^0 = g_j^0$. 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

$$\int_{\mathbb{R}} f_\zeta \left(w'\beta^0 + \alpha \right) f_{\alpha^0_{g_i^0}}(\alpha) d\lambda(\alpha) = \int_{\mathbb{R}} f_\zeta \left(w'\beta^0 + \alpha \right) f_\alpha(\alpha) d\lambda(\alpha) \quad \forall w \in \mathcal{X}^0, \quad (4.9.5)$$

verifies $\mathcal{S}^* = \left\{ f \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda) : f_\alpha = f_{\alpha^0_{g_i^0}} \text{ a.e.} \right\}$. Suppose $f^*_\alpha \in \mathcal{S}^*$ and consider the change of variable $z = w'\beta^0 + \alpha$ in (4.9.5). Then, for all $\delta \in (x'\beta^0 - \beta_1^0\epsilon, x'\beta^0 + \beta_1^0\epsilon) \subset \mathbb{R}$,

$$\int_{\mathbb{R}} f_\zeta(z) f_{-\alpha^0_{g_i^0}}(\delta - z) d\lambda(z) = \int_{\mathbb{R}} f_\zeta(z) f^*_{-\alpha}(\delta - z) d\lambda(z). \quad (4.9.6)$$

Note that both sides of Eq. (4.9.6) are convolutions of f_ζ with $df_{-\alpha^0_{g_i^0}}$ or $df^*_{-\alpha}$. By letting

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

and using commutativity of the convolution product, Eq. (4.9.6) implies that there exists an open set $U \subset \mathbb{R}$ such that

$$\mathcal{W}(\delta) = 0, \quad \forall \delta \in U. \quad (4.9.7)$$

Given Assumption 4.9.1(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. (4.9.7) holds for all $\delta \in \mathbb{R}$, i.e.:

$$\int_{\mathbb{R}} f_\zeta(\delta - z) \left[f_{-\alpha^0_{g_i^0}}(z) - f^*_{-\alpha}(z) \right] d\lambda(z) = 0, \quad \forall \delta \in \mathbb{R}. \quad (4.9.8)$$

Since the functions f_ζ, $f_{-\alpha^0_{g_i^0}}$, and $f^*_{-\alpha}$ belong to $L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$, I can apply Fourrier transformation on both sides of Eq. (4.9.8) to obtain

$$\varphi_{f_\zeta}(v) \times \left[\varphi_{f_{-\alpha^0_{g_i^0}}}(v) - \varphi_{f^*_{-\alpha}}(v) \right] = 0, \quad \forall v \in \mathbb{R}, \quad (4.9.9)$$

where φ_f is the Fourrier transform of f. By Assumption 4.9.1(c) again, the set

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

is of zero Lebesgue measure. Equation (4.9.9) therefore implies $\varphi_{f_{-\alpha^0_{g_i^0}}} = \psi_{f^*_{-\alpha}}$ a.e.. Since Fourrier transforms are continuous, I obtain $\varphi_{f_{-\alpha^0_{g_i^0}}} = \varphi_{f^*_{-\alpha}}$ everywhere and thus $f_{\alpha^0_{g_i^0}} = f^*_\alpha$ everywhere.
The proof of Lemma 4.9.1 is complete.

4.9.3 Proof of Theorem 4.5.1

The key argument is to linearize problem (4.5.2) 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 $\theta = (\beta', \alpha', \gamma')' \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0T}$, 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 $Z_{it}^0 = X_{it}'\beta^0 + \alpha_{g_i^0 t}^0$ denote Z_{it} evaluated at the true parameter value θ^0. Note that the NGFE estimator $\widehat{\theta}$ minimizes $\widehat{Q}(\cdot)$ over all $\theta \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0T}$. Define the auxiliary quadratic function:

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

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 4.5.2(a) and 4.5.2(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 4.5.1(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 $z_1, z_2 \in \mathcal{Z}$

$$- \ln \Psi(z_1) \geq - \ln \Psi(z_2) + s(z_2)(z_1 - z_2) + \frac{b_{\min}}{2} (z_1 - z_2)^2, \tag{4.9.10}$$

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 (4.9.10) over i, t, I have, for all $\theta \in \mathcal{B} \times \mathcal{G}^{0N} \times \mathcal{A}^{G^0T}$,

$$\widehat{Q}(\theta) - \widehat{Q}(\theta^0) \geq \frac{b_{\min}}{2} \check{Q}(\theta) + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E_{it} \left(Q_{it} \left(Z_{it} - Z_{it}^0 \right) \right), \tag{4.9.11}$$

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

$$0 \geq \widehat{Q}(\widehat{\theta}) - \widehat{Q}(\phi^0) \geq \frac{b_{\min}}{2} \check{Q}(\widehat{\theta}) + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} E_{it} \left(Q_{it} \left(\widehat{Z}_{it} - Z_{it}^0 \right) \right), \tag{4.9.12}$$

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 4.9.2 *Let Assumption 4.5.1 and Assumptions 4.5.2(a)-(b) hold. Then,*

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

Proof of Lemma 4.9.2: The proof closely follows that of Lemma A.1 in Bonhomme and Manresa (2015), up to a few adjustments.

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

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 4.5.1(a) and 4.5.1(b), for all $s<t$, I have

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

where the penultimate equality follows because $\Psi'(Z_{it}^0)$ is $\mathcal{F}_t$-measurable, and the last equality follows from $\mathbb{E}(Y_{it}|\mathcal{F}_t)=\Psi(Z_{it}^0)$. By Cauchy-Schwarz (CS) inequality, and using Assumption 4.5.1(b), 4.5.2(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

$$\left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left(Q_{it}Q_{is}E_{it}E_{is}X_{it}'X_{is}\right)\right|\leq M'. \tag{4.9.13}$$

By (4.9.13), 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, $\|\beta - \beta^0\|$ is bounded under Assumption 4.5.2(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 $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Q_{it}E_{it}\alpha_{g_it}$ 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

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

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

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

where, by Assumption 4.5.2(a), $\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}^2$ is uniformly bounded. Now, note that

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

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

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

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,

$$\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}(Q_{it}Q_{jt}E_{it}E_{jt}-\mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt}))\right|\right)^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}(Q_{it}Q_{jt}E_{it}E_{jt}-\mathbb{E}(Q_{it}Q_{jt}E_{it}E_{jt}))\right)^2. \tag{4.9.14}$$

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 (4.9.14) 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 4.9.2. □

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

$$\check{Q}(\widehat{\theta}) \geq \widehat{\rho}\left\|\widehat{\beta}-\beta^0\right\|^2, \tag{4.9.15}$$

where $\mathrm{plim}_{N,T\to\infty}\widehat{\rho} = \rho > 0$. Hence, combining (4.9.12), Lemma 4.9.2, and (4.9.15) 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

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

which is $o_p(1)$ by Assumptions 4.5.2(a)-4.5.2(b), by consistency of $\widehat{\beta}$, and because Lemma 4.9.2 and (4.9.12) together imply $\check{Q}(\widehat{\theta}) = o_p(1)$.
This completes the proof of Theorem 4.5.1.

4.9.4 Proof of Theorem 4.5.2

Step 1: A Useful Asymptotic Equivalence

Lemma 4.9.7 below provides an asymptotic equivalence result which is key to prove Theorem 4.5.2. I first prove three lemmas (4.9.3, 4.9.4, and 4.9.5) that help in showing that NGFE estimators achieve uniformly consistent classification of individuals (Lemma 4.9.6). This, in turn, allows me to prove Lemma 4.9.7.

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 $\mathbb{R}^{G^0T}$, 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 4.9.3 *Let Assumptions 4.5.1-4.5.2, and 4.5.3(a)-4.5.3(b) hold. Then, as N and T tend to infinity,*

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

Proof of Lemma 4.9.3: Given Theorem 4.5.1, 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 4.9.4 and Lemma 4.9.5 are direct consequences of Theorem 6.2 in Rio (2000) (see also Merlevède et al., 2011) and Theorem 3.2 in Lesigne and Volný (2001), respectively.

Lemma 4.9.4 (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 $\mathbb{P}(|z_t|<z) \leq \exp\left(1-(\frac{z}{b})^{d_2}\right)$, where a, b, d_1, and d_2 are positive constants. Then, for all $z>0$, for all $\delta>0$,*

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

Lemma 4.9.5 [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 $t=1,\ldots,T$. Then, for $a>0$, there exist positive constants A and B such that for all $z \geq a/\sqrt{T}$*

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

[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.

Proof of Lemma 4.9.5: In the proof of their Theorem 3.2 Lesigne and Volný (2001) show that if $E(\exp(|z_t|) \leq M$ for all $t = 1, \ldots, T$, then for any $x > 0$ and $t \in (0, 1)$, I have

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{t=1}^{T} z_t\right| > Tz\right) \\
&\quad < \left(2 + \frac{M}{(1-t)^2}\left[\frac{1}{4}t^{4/3}(z^{-2}T^{-1})^{1/3} + t^{2/3}(z^{-2}T^{-1})^{2/3} + 2z^{-2}T^{-1}\right]\right) \\
&\qquad \times \exp\left(-\frac{1}{2}t^{2/3}(z^2T)^{1/3}\right). \qquad (4.9.16)
\end{aligned}$$

Note that $\mathbb{P}\left(\left|\sum_{t=1}^{T} z_t\right| > Tz\right) = \mathbb{P}\left(\left|\sum_{t=1}^{T}(\delta z_t)\right| > T(\delta z)\right)$ 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 (4.9.16) yields

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

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 4.9.6 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

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(g)t} - \alpha^0_{gt}\right)^2 \xrightarrow{p} 0. \qquad (4.9.17)$$

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 $(\beta, \alpha) \in \mathcal{B} \times \mathcal{A}^{G^0 T}$ that satisfy $\|\beta - \beta^0\|^2 < \eta$ and $\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt} - \alpha^0_{gt}\right)^2 < \eta$ for all $g \in \mathcal{G}^0$.

Lemma 4.9.6 *For $\eta > 0$ small enough, I have, for all $\delta > 0$ and as N and T tend to infinity,*

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

Proof of Lemma 4.9.6: Note that, from the definition of $\widehat{g}_i(\cdot)$, for all $g \in \mathcal{G}^0$,

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

so

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\widehat{g}_i(\beta,\alpha)\neq g_i^0\}=\sum_{g\in\mathcal{G}^0}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{g_i^0\neq g\}\mathbb{1}\{\widehat{g}_i(\beta,\alpha)=g\}$$
$$\leq\sum_{g\in\mathcal{G}^0}\frac{1}{N}\sum_{i=1}^{N}W_{ig}(\beta,\alpha),$$

where

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

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 4.5.1(b), and 4.5.2(a)-4.5.2(b), $v\mapsto\ln\left(\Psi\left(Q_{it}\left(X_{it}'v+\alpha_{gt}\right)\right)\right)$ is uniformly Lipschitz over $(i,t,\alpha,g)\in\{1,\ldots,N\}\times\{1,\ldots,T\}\times A^{G^0T}\times\mathcal{G}^0$, i.e., there exists a constant $L_\beta>0$ such that, for all $(i,t,\alpha,g)\in\{1,\ldots,N\}\times\{1,\ldots,T\}\times A^{G^0T}\times\mathcal{G}^0$, all $\beta_1,\beta_2\in\mathcal{B}$, almost surely

$$\left|\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta_1+\alpha_{gt}\right)\right)\right)-\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta_2+\alpha_{gt}\right)\right)\right)\right|\leq L_\beta\left\|\beta_1-\beta_2\right\|. \quad (4.9.18)$$

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

$$\left|\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+a\right)\right)\right)-\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta+b\right)\right)\right)\right|\leq L_\alpha\left|a-b\right|. \quad (4.9.19)$$

Then, by choosing $g=g_i^0,\beta_1=\beta^0$ and $\beta_2=\beta$ in (4.9.18), I have, for all (β,α) and all i,

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

By choosing $a=\alpha_{g_i^0t}$, $b=\alpha_{g_i^0}^0$, and $\beta=\beta^0$ in (4.9.19), I have, for all (β,α) and all i,

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

where I used the norm inequality $\|u\|_1\leq\sqrt{T}\|u\|\leq T\|u\|$ for all $u\in\mathbb{R}^T$, $T\in\mathbb{N}^*$, where $\|\cdot\|_1$ 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 (4.9.3), yields

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

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

As I have

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

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

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

where C_1, C_2, C_3,

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

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

$$W_{ig}(\beta,\alpha)\leq\max_{\widetilde{g}\neq g}\mathbb{1}\Big\{\sum_{t=1}^{T}V_{it}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)\leq-\frac{b_{\min}}{2}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2$$
$$+T\sqrt{\eta}\left[(c_1\vee c_2)\times(M+C_1)+b_{\min}C_2M+C_3+L_\beta+L_\alpha\right]\Big\}.$$

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

$$\overline{W}_{ig} = \max_{\widetilde{g}\neq g} \mathbb{1}\Big\{ \sum_{t=1}^{T} V_{it} \left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{b_{\min}}{2} \sum_{t=1}^{T} \left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \tag{4.9.20}$$

$$+ T\sqrt{\eta}\left[(c_1 \vee c_2) \times (M + C_1) + b_{\min} C_2 M + C_3 + L_\beta + L_\alpha\right] \Big\}. \tag{4.9.21}$$

As a result,

$$\sup_{(\beta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i(\beta,\alpha) \neq g_i^0\} \leq \frac{1}{N} \sum_{i=1}^{N} \sum_{g\in\mathcal{G}^0} \overline{W}_{ig}. \tag{4.9.22}$$

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

$$\begin{aligned}
\mathbb{P}\left(\overline{W}_{ig} = 1\right) &\leq \sum_{\widetilde{g}\neq g} \mathbb{P}\Bigg(\sum_{t=1}^{T} V_{it} \left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -\frac{b_{\min}}{2} \sum_{t=1}^{T} \left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \\
&\qquad + T\sqrt{\eta}\left[(c_1 \vee c_2) \times (M + C_1) + b_{\min} C_2 M + C_3 + L_\beta + L_\alpha\right] \Bigg) \\
&\leq \sum_{\widetilde{g}\neq g} \Bigg\{ \mathbb{P}\left(\frac{1}{T} \sum_{t=1}^{T} \left(\alpha^0_{gt} - \alpha^0_{\widetilde{g}t}\right)^2 \leq \frac{c_{g,\widetilde{g}}}{2} \right) \\
&\qquad + \mathbb{P}\Bigg(\sum_{t=1}^{T} V_{it} \left(\alpha^0_{\widetilde{g}t} - \alpha^0_{gt}\right) \leq -T\frac{c_{g,\widetilde{g}} b_{\min}}{4} \\
&\qquad + T\sqrt{\eta}\left[(c_1 \vee c_2) \times (M + C_1) + b_{\min} C_2 M + C_3 + L_\beta + L_\alpha\right] \Bigg) \Bigg\}.
\end{aligned} \tag{4.9.23}$$

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 4.5.1(a) and 4.5.1(b) since

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

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

$$\mathbb{E}\left(Y_{it}|X_{i1},\ldots,X_{it},\alpha^0,\gamma^0\right) - \Psi\left(Z_{it}^0\right) = 0.$$

By Assumption 4.5.2(b), for all $i \in \{1,\ldots,N\}$, $\left\{V_{it}\left(\alpha_{\widetilde{gt}}^0 - \alpha_{gt}^0\right) : t\right\}$ is such that $\left|V_{it}\left(\alpha_{\widetilde{gt}}^0 - \alpha_{gt}^0\right)\right| \leq (\widetilde{c}_1 \vee \widetilde{c}_2) < \infty$, 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 4.9.5, there exist positive constants A and B, independent from (i,t), such that for all $z > a/\sqrt{T}$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} V_{it}\left(\alpha_{\widetilde{gt}}^0 - \alpha_{gt}^0\right)\right| \geq z\right) \leq A\exp\left(-B(z^2T)^{1/3}\right). \tag{4.9.24}$$

I now bound the two terms on the right-hand side of (4.9.23).

- By applying Lemma 4.9.4, and conducting the same reasoning as in the first bullet point page 1176 in Bonhomme and Manresa (2015), under Assumptions 4.5.2(a) and 4.5.3(b)-(c), for all $\delta > 0$ and as T tends to infinity,

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

- Lastly, to bound the second term on the right-hand side of (4.9.23), I denote as $\underline{c}$ the minimum of $c_{g,\widetilde{g}}$ over all $g \neq \widetilde{g}$ and I take

$$\eta \leq \left(\frac{\underline{c}}{8[(c_1 \vee c_2) \times (M + C_1) + b_{\min}C_2M + C_3 + L_\beta + L_\alpha]}\right)^2. \tag{4.9.25}$$

  Note that this upper bound on η does not depend on T. Taking η satisfying (4.9.25) yields, for all $\widetilde{g} \neq g$,

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

  Lastly, by applying (4.9.24) with $z = \frac{c_{g,\widetilde{g}}}{8}$, for T sufficiently large, I obtain

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

  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 (4.9.23), that for η satisfying (4.9.25) and for all $\delta > 0$,

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{g\in\mathcal{G}^0}\mathbb{P}\left(\overline{W}_{ig}=1\right) \leq \left|\mathcal{G}^0\right|\left(\left|\mathcal{G}^0\right|-1\right)\left[o(T^{-\delta})+o(T^{-\delta})\right] = o(T^{-\delta}). \quad (4.9.26)$$

To complete the proof of Lemma 4.9.6, note that, for η that satisfies (4.9.25), I have, for all $\delta > 0$ and all $\varepsilon > 0$,

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

where I have used (4.9.22), the Markov inequality, and (4.9.26), respectively. This ends the proof of Lemma 4.9.6. □

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

Lemma 4.9.7 (Asymptotic Equivalence) *Let Assumptions 4.5.1, 4.5.2, and 4.5.3 hold. Then, for all $\delta > 0$ and as N and T tend to infinity*

$$\mathbb{P}\left(\sup_{i\in\{1,\ldots,N\}}\left|\widehat{g}_i - g_i^0\right| > 0\right) = o(1) + o(NT^{-\delta}), \quad (4.9.27)$$

and

$$\widehat{\beta} = \widetilde{\beta} + o_p(T^{-\delta}), \quad (4.9.28)$$

and

$$\widehat{\alpha}_{gt} = \widetilde{\alpha}_{gt} + o_p(T^{-\delta}) \text{ for all } g, t. \quad (4.9.29)$$

Proof of Lemma 4.9.7: The proof closely follows pages 1178-1180 in Bonhomme and Manresa (2015).

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

$$\widehat{Q}(\beta,\alpha) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{\widehat{g}_i(\beta,\alpha)t}\right)\right)\right), \quad (4.9.30)$$

$$\widetilde{Q}(\beta,\alpha) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} -\ln\left(\Psi\left(Q_{it}\left(X_{it}'\beta + \alpha_{g_i^0 t}\right)\right)\right). \quad (4.9.31)$$

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 4.9.6 holds. Using Assumptions 4.5.2(a) and 4.5.2(b), it is then easy to see that, for all $\delta > 0$,

$$\sup_{(\beta,\alpha)\in\mathcal{N}_\eta}\left|\widehat{Q}(\beta,\alpha) - \widetilde{Q}(\beta,\alpha)\right| = o_p(T^{-\delta}). \quad (4.9.32)$$

By consistency of $\widehat{\beta}$ (Theorem 4.5.1) and $\widehat{\alpha}$ (Lemma 4.9.3), and because $\widetilde{\beta}$ and $\widetilde{\alpha}$ are also consistent under the conditions of Theorem 4.5.1, I have, as N and T tend to infinity,

$$\mathbb{P}\left(\left(\widehat{\beta},\widehat{\alpha}\right)\notin\mathcal{N}_\eta\right)\to 0, \tag{4.9.33}$$

$$\mathbb{P}\left(\left(\widetilde{\beta},\widetilde{\alpha}\right)\notin\mathcal{N}_\eta\right)\to 0. \tag{4.9.34}$$

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. (4.9.32)-(4.9.34) imply

$$\widetilde{Q}(\widehat{\beta},\widehat{\alpha})-\widetilde{Q}(\widetilde{\beta},\widetilde{\alpha})=o_p(T^{-\delta}). \tag{4.9.35}$$

Now, using that $(\widetilde{\beta},\widetilde{\alpha})$ minimizes the twicely continuously differentiable function $\widetilde{Q}(\cdot)$, I obtain under Assumption 4.5.1(b)

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

where $\widehat{\rho}\xrightarrow{p}\rho>0$ as a consequence of Assumption 4.5.2(c). Hence, $\widetilde{\beta}-\widehat{\beta}=o_p(T^{-\delta})$ for all $\delta>0$. This shows (4.9.28).

#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 4.9.7 is complete. □

Step 2: Asympotic Properties of the Oracle MLE

By Lemma 4.9.7 and Slutsky's lemma, it is sufficient to analyze the limiting distribution of the unfeasible maximum likelihood estimator, $(\widetilde{\beta},\widetilde{\alpha})$, defined as

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

where

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

First, I show

$$\sqrt{NT}\left(\widetilde{\beta}-\beta^0\right)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_\beta^{-1}\right). \tag{4.9.36}$$

Second, I show for all g, t,

$$\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{gt}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\widetilde{\pi}^2_{gt}}\right), \tag{4.9.37}$$

and conclude by Slutsky's lemma.

 # 1. (4.9.36) **holds.** Under Assumption 4.5.4, 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 $B \in \mathbb{R}^p$ and $S_{NT} \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\beta^{-1}\right)$. The result then follows from $T = o(N)$.

#2. (4.9.37) **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

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

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

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} Q_{it}\left(\ln\Psi\right)'\left(Q_{it}\left(X_{it}'\beta + \widetilde{\alpha}_{gt}(\beta)\right)\right) = 0. \tag{4.9.38}$$

Differentiating Eq. (4.9.38) with respect to β yields

$$\frac{\mathrm{d}\widetilde{\alpha}_{gt}(\beta)}{\mathrm{d}\beta} = -\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\}\left(\ln\Psi_{it}\right)''\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\}\left(\ln\Psi_{it}\right)'' X_{jt}\right), \tag{4.9.39}$$

where $\left(\ln\Psi_{it}\right)'' := \left(\ln\Psi\right)''\left(Q_{it}\left(X_{it}'\beta + \widetilde{\alpha}_{g_i^0 t}(\beta)\right)\right)$. By Taylor's theory, Eq. (4.9.39) and Assumptions 4.5.2(a)-(b) imply that there exists $C > 0$ such that, almost surely,

$$\sup_{\beta,\beta' \in = \mathcal{B}} \left|\widetilde{\alpha}_{gt}(\beta) - \widetilde{\alpha}_{gt}(\beta')\right| \leq C\left\|\beta - \beta'\right\|. \tag{4.9.40}$$

Deduce that

$$\begin{aligned}
\sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha^0_{gt}\right) &= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha^0_{gt}\right) + \sqrt{N}\left(\widetilde{\alpha}_{gt}(\widetilde{\beta}) - \widetilde{\alpha}_{gt}(\beta^0)\right) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha^0_{gt}\right) + O_p\left(\sqrt{N}\left\|\widetilde{\beta} - \beta^0\right\|\right) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha^0_{gt}\right) + O_p(1/\sqrt{T}) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt}(\beta^0) - \alpha^0_{gt}\right) + o_p(1),
\end{aligned} \tag{4.9.41}$$

where the second and third equality use Eq. (4.9.40) and (4.9.36) respectively. Now, by expanding each summand in Eq. (4.9.38) 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

$$\widetilde{\alpha}_{gt}(\beta^0) = \alpha_{gt}^0 - \left(\sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \left(-\ln \Psi\right)'' \left(Q_{it} Z_{it}^*\right)\right)^{-1} \left(\sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} Q_{it} \left(-\ln \Psi\right)' \left(Q_{it} Z_{it}^0\right)\right). \tag{4.9.42}$$

Equation (4.9.42) yields

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

where the second equality follows from $\sup_{i=1,\ldots,N} \left|Z_{it}^* - Z_{it}^0\right| = o_p(1)$ (it is easy to prove that $\widetilde{\alpha}_{gt}(\beta^0) - \alpha_{gt}^0 = o_p(1)$ using (4.9.42), Assumptions 4.5.1(b), 4.5.2(a)-(b), and 4.5.4(e)) and Assumption 4.5.4(c), and the last convergence follows by Assumption 4.5.4(e). Given (4.9.41), (4.9.37) follows by Slutsky's lemma.

#3. Conclusion. Let $\delta > 0$. By Lemma 4.9.7,

$$\begin{aligned}
\sqrt{NT}\left(\widehat{\beta} - \beta^0\right) &= \sqrt{NT}\left(\widetilde{\beta} - \beta^0\right) + \sqrt{NT}\left(\widehat{\beta} - \widetilde{\beta}\right) \\
&= \sqrt{NT}\left(\widetilde{\beta} - \beta^0\right) + o_p\left(\sqrt{N}T^{1-\delta}\right),
\end{aligned} \tag{4.9.43}$$

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

$$\begin{aligned}
\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha_{gt}^0\right) &= \sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha_{gt}^0\right) + \sqrt{N}\left(\widehat{\alpha}_{gt} - \widetilde{\alpha}_{gt}\right) \\
&= \sqrt{N}\left(\widetilde{\alpha}_{gt} - \alpha_{gt}^0\right) + o_p\left(\sqrt{N}T^{-\delta}\right).
\end{aligned} \tag{4.9.44}$$

Since (4.9.43) and (4.9.44) 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

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

This result, combined with (4.9.36), (4.9.37), and Slustky's lemma yields (4.5.4) and (4.5.5).

4.10 Extensions

4.10.1 Cluster-Specific Slopes and Time-Specific Effects

In this section, I consider the following extension of model (4.2.1): for all $(i,t) \in \mathcal{N} \times \mathcal{T}$,

$$\mathbb{P}\left(Y_{it} = y | X_{i1}, \ldots, X_{it}, \alpha^0_{g^0_i t}, \beta^0_{g^0_i}, g^0_i, \zeta^0_t\right) = h^0\left(y, X'_{it}\beta^0_{g^0_i} + \alpha^0_{g^0_i t} + \zeta^0_t\right), \qquad (4.10.1)$$

where $h^0 \in \mathcal{H}$, $\|\beta^0_1\| = 1$ 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 $\{(i,t) \in \mathcal{N} \times \mathcal{T} : g^0_i = g\}_{g \in \mathcal{G}^0}$. Here, the difficulty arises from the assumption that the group membership variables g^0_i are unknown to the econometrician. Let $\beta^0 := \{\beta^0_g : g\}$. I first adapt Assumption 4.3.1:

Assumption 4.10.1 (Random sampling)

(a) $(Y'_i, X'_i, g^0_i)'$ *is i.i.d. across* $i \in \mathcal{N}$ *conditional on* α^0, β^0, λ^0, μ^0.

(b) For all $i \in \mathcal{N}$: $\{(Y_{it}, X'_{it}, \alpha^0_{g^0_i t}, \zeta^0_t)'\}_{t \geq 2}$ *is a strictly stationary strong mixing process with mixing coefficients* $\tau_i(\cdot)$ *conditional on* $g^0_i, \mu^0_{g^0_i}, \xi^0_i, \beta^0_{g^0_i}$. *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 $t \in \mathcal{T}$: $Y_{1t}|X_{1t}, g^0_1, \alpha^0, \beta^0, \lambda^0, \mu^0, \xi^0 \overset{d}{=} Y_{1t}|X_{1t}, g^0_1, \alpha^0_{g^0_1 t}, \beta^0_{g^0_i}$.

Assumption 4.10.2 (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* $\rho_i(x) : \mathcal{X}^0 \ni x \mapsto \mathbb{P}\left(Y_{i2} = y | X_{i2} = x, \beta^0_{g^0_i}, g^0_i, \mu^0_{g^0_i}, \xi^0_i\right)$, $\phi(\rho_i, \rho_j) = \mathbb{1}\{g^0_i = g^0_j\}$.

(b) For all $g \in \mathcal{G}^0$, *almost surely* $\mathbb{P}(g^0_1 = g | \alpha^0, \beta^0, \lambda^0, \mu^0, \xi^0) > 0$.

Assumption 4.10.3 (Regularity and smoothness)

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

(b) Almost surely, $\mathbb{E}(\|X_{12}\|^2 | g^0_1, \alpha^0, \beta^0, \lambda^0, \mu^0)$ *is finite and* $\mathbb{E}(X_{12}X'_{12}|g^0_1, \alpha^0, \beta^0, \lambda^0, \mu^0)$ *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* $X'_{it}\beta^0_{g^0_i} + \alpha^0_{g^0_i t}$.

Assumption 4.10.4 (Monotonicity) *There exists* $y \in \mathcal{Y}$ *such that* $h^0(y, v)$ *is strictly monotonic in* v.

Assumption 4.10.5 (Compensating variations)

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

$$\alpha_{gt}^0 + x_1'\beta_g^0 + \zeta_t^0 = \alpha_{g\widetilde{t}}^0 + x_2'\beta_g^0 + \zeta_{\widetilde{t}}^0. \tag{4.10.2}$$

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

$$\alpha_{gt}^0 + x_3'\beta_g^0 + \zeta_t^0 = \alpha_{\widetilde{g}t}^0 + x_4'\beta_{\widetilde{g}}^0 + \zeta_t^0. \tag{4.10.3}$$

Theorem 4.10.1 (Identification) *Let Assumptions 4.10.1, 4.10.2 and 4.10.3(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 4.10.3(b)-4.10.5 further hold, then*
   - *β^0 is identified.*
   - *$\zeta_t^0 + \alpha_{gt}^0$ is identified for all $(g, t) \in \mathcal{G}^0 \times \mathbb{N}^*$.*

Proof of Theorem 4.10.1: The proofs of Part 1 and identification of β^0 are identical to the corresponding parts of the proof of Theorem 4.3.1, up to running nonparametric regressions for all $g \in \mathcal{G}^0$ to identify β_g^0. Next, Assumption 4.10.5(b) ensures that, for all $(g, \widetilde{g}, t)$, I can identify $(x_1, x_2) \in \mathcal{X}^2$, such that for some $y \in \mathcal{Y}$,

$$h^0\left(y, x_1'\beta_g^0 + \alpha_{gt}^0 + \zeta_t^0\right) = h^0\left(y, x_2'\beta_{\widetilde{g}}^0 + \alpha_{\widetilde{g}t}^0 + \zeta_t^0\right).$$

By inverting $h^0(y, \cdot)$, I obtain $\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0 = x_1'\beta_g^0 - x_2'\beta_{\widetilde{g}}^0$. Since the right-hand side is identified, $\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0$ is identified for all $(g, \widetilde{g}, t)$. In particular, $(\alpha_{g1}^0)_{g \in \mathcal{G}^0}$ is identified. Now, suppose that $G^0 \geq 2$. By Assumption 4.10.5(a), for all $(g, t, \widetilde{t})$, I can identify $(x_3, x_4) \in \mathcal{X}^2$ such that, for some $y \in \mathcal{Y}$,

$$h^0\left(y, x_3'\beta_g^0 + \alpha_{gt}^0 + \zeta_t^0\right) = h^0\left(y, x_4'\beta_g^0 + \alpha_{g\widetilde{t}}^0 + \zeta_{\widetilde{t}}^0\right). \tag{4.10.4}$$

By inverting $h^0(y, \cdot)$ again, Eq. (4.10.4) yields

$$\zeta_t^0 - \zeta_{\widetilde{t}}^0 = \alpha_{g\widetilde{t}}^0 - \alpha_{gt}^0 + (x_4 - x_3)'\beta_g^0. \tag{4.10.5}$$

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).

4.10.2 Group and Time-Specific Link Functions

Consider the general model:

$$\mathbb{P}\left(Y_{it} = y | X_i^t, g_i^0\right) = h_t^0\left(y, X_{it}'\beta^0, g_i^0\right), \quad i = 1, \ldots, N, t = 1, \ldots, T. \tag{4.10.6}$$

Under an adaptation of Assumption 4.3.2, the same analysis can be conducted to identify g_i^0 and $(h_t^0)_{t\geq 1}$ up to group relabeling, and β^0 up to scale.

4.10.3 Grouping Time Periods

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

$$\mathbb{P}\left(Y_{it} = y | X_i^t, \alpha^0_{g_i^0 k_t^0}, g_i^0, k_t^0\right) = h^0\left(y, X_{it}'\beta^0 + \alpha^0_{g_i^0 k_t^0}\right), \quad i = 1,\dots,N, t = 1,\dots,T \tag{4.10.7}$$

When $\mathcal{N} = \mathcal{T}$, this gives rise to a so-called Holland et al. (1983)'s stochastic block model on latent variables. Methods developed in Chapters 3 and in 4 can be used to obtain identification results for nonlinear multiplicative models in cases where $G^0 = K^0$ and under symmetry ($\alpha^0_{\underset{\sim}{gg}} = \alpha^0_{\underset{\sim}{gg}}$ almost surely).

4.10.4 NGFE Large Sample Theory for Poisson Count Models

Theorem 4.5.1 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 4.10.6

(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

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

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 4.10.6(b) is verified by the Poisson count model (9).

Theorem 4.10.2 (Consistency in General Nonlinear Models) *Let Assumptions 4.5.2 and 4.10.6 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 4.5.2 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. (4.9.19) and (4.9.18) 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 $\exp(Z_{it}^0)$,

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

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

$$\widehat{\pi}_{gt} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{g_i^0 = g\} \exp(Z_{it}^0).$$

Consider the following assumption.

Assumption 4.10.7

(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*

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

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

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

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

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

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

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

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

(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* $\mathcal{A}^{G^0 T}$*.*

Theorem 4.10.3 (Asymptotic Distribution in the Poisson Count Model – Conjectured) *Let Eq.* (4.2.3)*, Assumptions 4.5.2, 4.5.3, and 4.10.7 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:

$$\sqrt{NT}\left(\widehat{\beta} - \beta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\beta}^{-1}\right), \tag{4.10.8}$$

and, for all (g,t),

$$\sqrt{N}\left(\widehat{\alpha}_{gt} - \alpha_{gt}^0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\widetilde{\pi}_{gt}^2}\right), \tag{4.10.9}$$

where Σ_{β}, ω_{gt}, and $\widetilde{\pi}_g$ are defined in Assumption 4.10.7.

4.11 Large-N, Large-T Inference

4.11.1 Binary Choice Model

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

$$\operatorname{Var}\left(\widehat{\alpha}_{gt}\right) = \frac{\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i = g\}\left((\ln\Psi)'\left(Q_{it}\left(X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}\right)\right)\right)^2}{\left(\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i = g\}\left(-\ln\Psi\right)''\left(Q_{it}\left(X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}\right)\right)\right)^2}. \tag{4.11.1}$$

Given Theorem 4.5.2, an estimate of the asymptotic variance of $\widehat{\beta}$ is

$$\operatorname{Var}\left(\widehat{\beta}\right) = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(-\ln\Psi\right)''\left(Q_{it}\left(X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}\right)\right)\left[X_{it} - \widehat{\widetilde{X}}_{\widehat{g}_i,t}\right]\left[X_{it} - \widehat{\widetilde{X}}_{\widehat{g}_i,t}\right]'\right)^{-1}, \tag{4.11.2}$$

where

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

4.11.2 Poisson Count Model

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

$$\operatorname{Var}\left(\widehat{\alpha}_{gt}\right) = \frac{\sum_{i=1}^{N}\mathbb{1}\{\widehat{g}_i = g\}\left(Y_{it} - \exp\left(X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}\right)\right)^2}{\left(\sum_{i=1}^{N}\mathbb{1}\{\widehat{g}_i = g\}\exp\left(X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}\right)\right)^2}. \tag{4.11.3}$$

Given Theorem 4.10.3, an estimate of the asymptotic variance of $\widehat{\beta}$ is

$$\operatorname{Var}\left(\widehat{\beta}\right) = \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\exp\left(X_{it}'\widehat{\beta} + \widehat{\alpha}_{\widehat{g}_i t}\right)\left[X_{it} - \widehat{\widetilde{X}}_{\widehat{g}_i,t}\right]\left[X_{it} - \widehat{\widetilde{X}}_{\widehat{g}_i,t}\right]'\right)^{-1}, \tag{4.11.4}$$

where

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

4.12 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

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

where

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

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 $(\theta_{\text{init}}', \alpha_{11\text{init}}, \ldots, \alpha_{G^0 T\text{init}})'$ 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 $w \overset{iid}{\sim} \mathcal{N}(0, (1/4)^2)$.

[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.

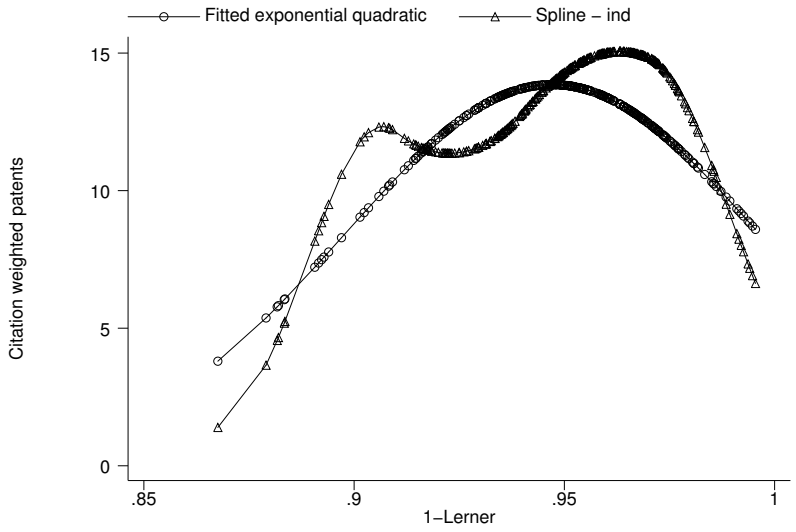
Computation Having large N is not computationally demanding. When T is very large, computation of the NGFE estimate might be demanding. The methods developed in Mugnier (2022) could be adapted. The statistical asymptotic results are confirmed by increasing (N, T) in unreported simulations.

4.13 Tables & Figures

4.13.1 Monte Carlo Simulations

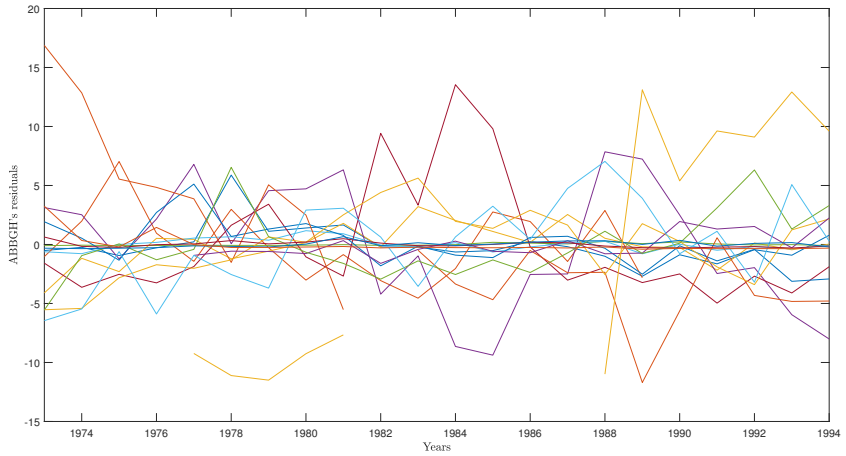
4.13.2 Empirical Application

FIGURE 4.1: Replicating Aghion et al. (2005)



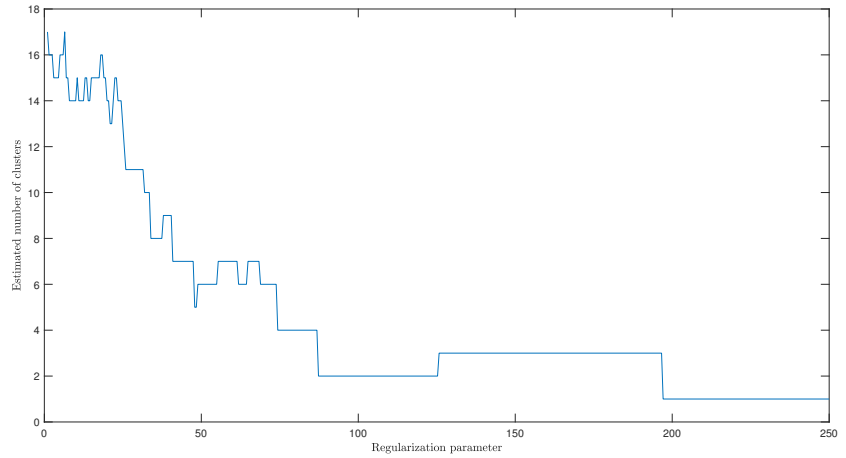
Notes: This figure replicates Aghion et al. (2005)'s Figure II. Data include 17 industries of 311 firms listed on the London Stock Exchange observed between 1973 − 1994. For each industry i at year t, the prediction replaces $\widehat{\nu}_i + \widehat{\xi}_t$ with an estimated constant $\widehat{\alpha}$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

FIGURE 4.2: Residuals of the Two-Way Fixed Effects Poisson Model



Notes: Each color represents an industry in Aghion et al. (2005)'s dataset. There are 17 industries observed over the period 1973-1994.

FIGURE 4.3: 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.

Table 4.1: 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 4.2: 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

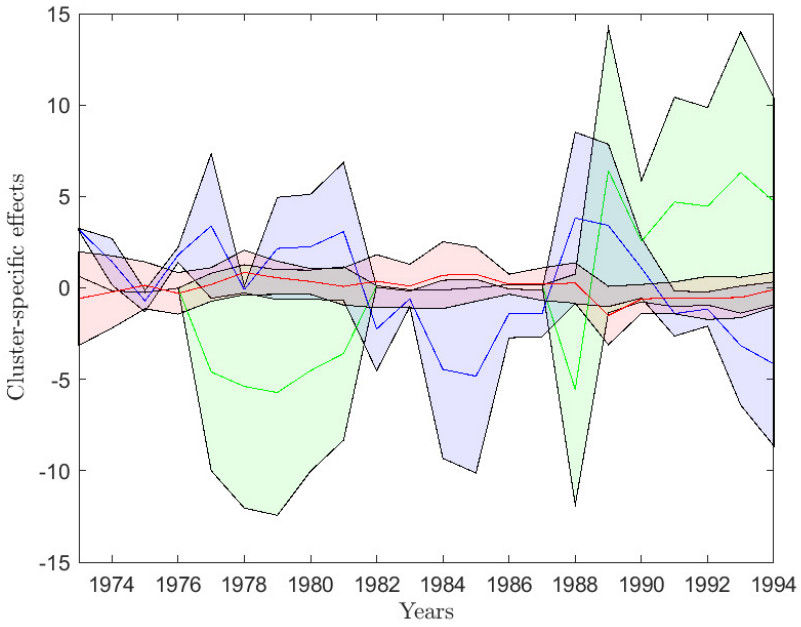
Notes: Static logit model with $\beta = 1$, $N = 90$, and $T = 7$. G^0 = true number of groups, P = Precision rate, R = Recall rate, RI = Rand Index, M = Misclassification Rate = minimum of $\sum_{i=1}^{N} \mathbb{1}\{\widehat{g}_i \neq g_i^0\}/N$ over all possible cluster relabelings, CPU = CPU time 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 4.3: 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 (4.11.4) 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.

FIGURE 4.4: 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.

Table 4.4: 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 4.1 for details.

Table 4.5: 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 4.2 for details.

TABLE 4.6: 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 4.3 for more details.

Table 4.7: 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

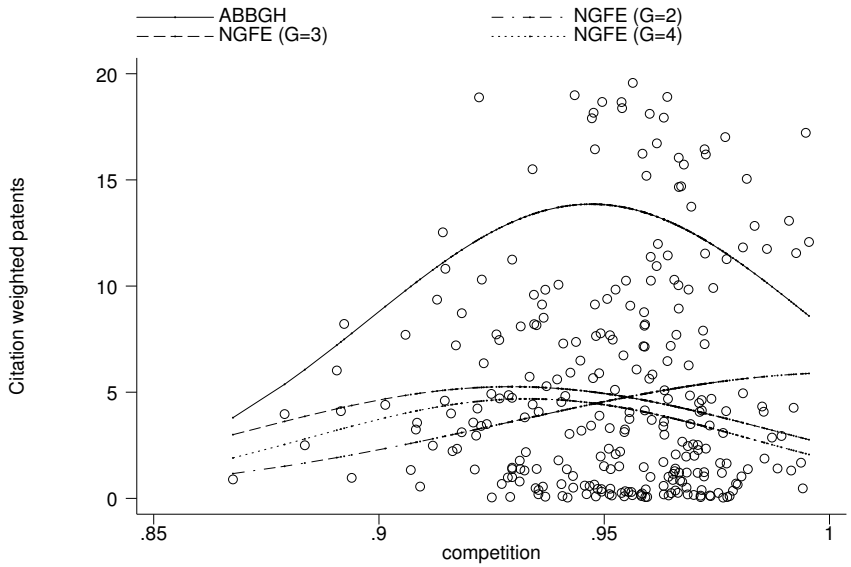
Notes: There are 17 industries and 354 observations over the time period 1973-94. See Aghion et al. (2005) for the exact definition of each variable.

Table 4.8: 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 4.5: Innovation and Competition Revisited: A Mildly Inverted-U Relationship



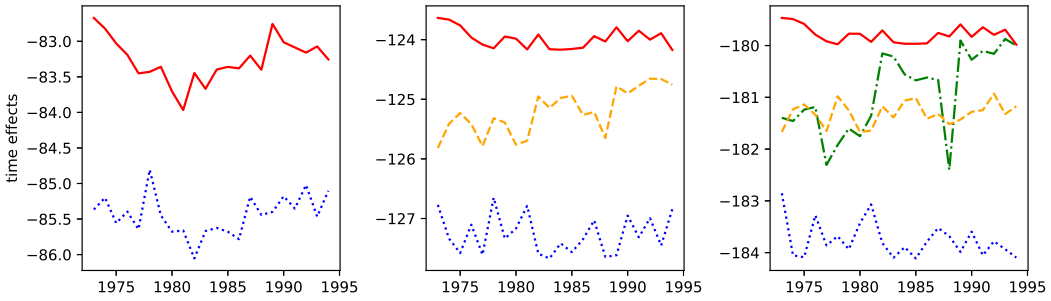
Notes: ABBGH (spe. (2) in Table 4.9) includes a constant and drop a time and an industry dummy (not included in the fit). NGFE (spe. (3), (4), and (5) in Table 4.9) does not specify a constant and averages the unobserved effects to obtain the intercept in the fit.

TABLE 4.9: 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

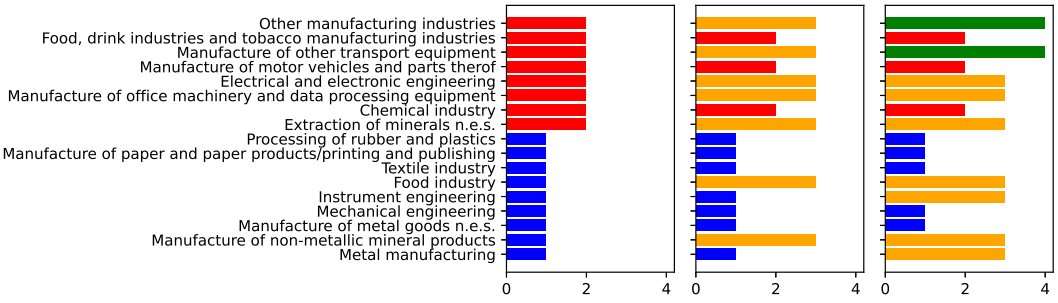
Notes: Analytical standard errors are under parentheses. The sample includes 354 observations from an unbalanced panel of 17 industries over the period 1973-1994. Competition$_{it}$ is measured by (1-Lerner index)$_{it}$ in the industry-year. NGFE estimates are computed using Lloyd's algorithm with 2,000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

FIGURE 4.6: Estimated Cluster-Specific Time-Varying Effects



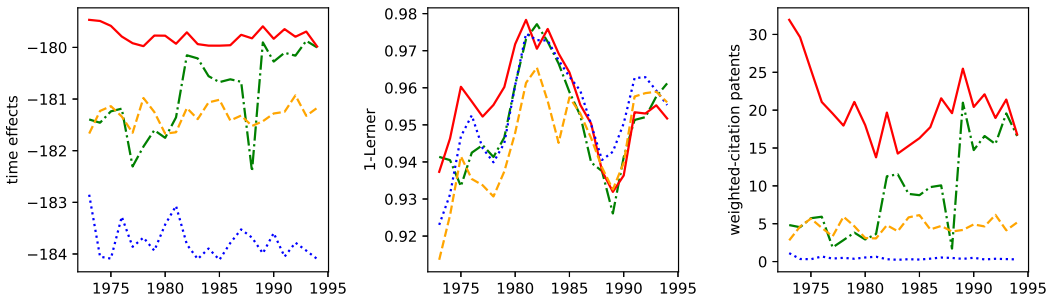
Notes: Solid red line = High-Innovation, dotted blue line = Low-Innovation, dashed orange line = Steady-Catchers, dashdotted green line = Noisy-Catchers. See Table 4.9 for more details.

FIGURE 4.7: 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 4.8: 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 4.10: 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.

Chapter 5

Asymptotic Properties of Empirical Quantile-Based Estimators

Voilà de ces distinctions savantes et délicates qui échappent à beaucoup de gens, parce que fort peu de gens réfléchissent; mais elles seront accueillies des gens instruits à qui je les adresse, et elles influeront, je l'espère, sur le nouveau Code que l'on nous prépare.

Le Marquis de Sade, *La Philosophie dans le boudoir*

Abstract: Developing statistical methods to infer the causal impact of policies on a given outcome based on non-experimental data is a workhorse challenge in econom(etr)ics. In a nonlinear extension of the popular Difference-in-Difference (DiD) framework, the parameter of interest, the average treatment effect (ATE), can be expressed as the difference between the expectation of some random variable (the outcome of the treated after treatment) and the expectation of an (unknown) quantile-CDF transform of another random variable (the outcome of the treated before treatment). A Changes-in-Changes estimator can be obtained by replacing expectations, quantile and CDF transforms by their empirical counterparts. In this chapter, I present new results showing that the asymptotic normality of such "Empirical Quantile-based" estimators holds under much weaker conditions that what is currently known. The proofs rely in particular on results on the standard empirical process and the theory of L-statistics. Finally, the finite sample behavior of the estimator is investigated through Monte Carlo simulations.[1]

5.1 Introduction

For any increasing function F on the real line, we denote by F^{-1} its left-continuous generalized inverse, $F^{-1}(q) = \inf\{x \in \mathbb{R} : F(x) \geq q\}$ for $q \in (0, 1]$, extended on $[0, 1]$

[1]This chapter is co-authored with Xavier D'Haultfœuille (CREST-ENSAE) and Jérémy L'Hour (Capital Fund Management).

by defining $F^{-1}(0) = 0$ when $\text{dom}(F) = [0,1]$. In particular, for any real-valued random variable W with cumulative distribution function (cdf) F_W, F_W^{-1} is the left-continuous quantile function. The associated empirical quantile function, $\widehat{F}_W^{-1}$, is obtained by replacing F_W by the standard empirical cdf obtained from a random sample $(W_i)_{i=1,\dots,n}$ in the definition of F_W^{-1}. All proofs are gathered in the appendix.

5.2 Asymptotic Results (Observed Rank)

We consider $\theta_0 = \mathbb{E}[F_Y^{-1}(U)]$ for some random variable U and the estimator

$$\widehat{\theta}_{n_1,n_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} \widehat{F}_Y^{-1}(U_j), \tag{5.2.1}$$

where $\widehat{F}_Y^{-1}$ is the empirical quantile function obtained from $(Y_i)_{i=1,\dots,n_1}$.

Assumption 5.2.1 (Sampling) *(i) $(Y_i)_{i=1,\dots,n_1}$ are independent draws from the distribution F_Y and $(U_j)_{j=1,\dots,n_2}$ are independent draws from the distribution F_U. (ii) U_j is independent of Y_i for any i and j. (iii) F_Y is continuous.*

Notice that (5.2.1) is a L-statistics (Shorack and Wellner, 1986, Chapter 19), i.e., $\widehat{\theta}_{n_1,n_2} = n_2^{-1}\sum_{i=1}^{n_2} c_{n_1 i} Y_{(i)}$, where $Y_{(1)} < \cdots < Y_{(n_1)}$ are the order statistics and $c_{n_1 i} = \#\{U_j : U_j \in ((i-1)/n_1, i/n_1]\}$. However, contrary to the textbook case, the weights c_{ni} are random variables,

$$(c_{n_1 1}, \dots, c_{n_1 n_1}) \sim \mathcal{M}(n_1, F_U(1/n_1) - F_U(0/n_1), \dots, F_U(n_1/n_1) - F_U((n_1-1)/n_1)).$$

In order to study its asymptotic behavior, let us decompose it into two parts, each only depending at the first order on the random sample $(Y_i)_{i=1,\dots,n_1}$ or $(U_i)_{i=1,\dots,n_2}$ but not both. Notice that θ_0 can be written as

$$\theta_0 = \int_0^1 F_Y^{-1}\, dF_U.$$

Let $\xi_i := F_Y(Y_i) \sim \mathcal{U}([0,1])$ and let $\mathbb{G}_{n_1}$ denote the empirical cumulative distribution function obtained from $(\xi_i)_{i=1,\dots,n_1}$. Thus, $\mathbb{G}_{n_1}^{-1}(\tau)$ is the usual empirical quantile of order τ.

The estimator (5.2.1) can be expressed as[2]

$$\widehat{\theta}_{n_1,n_2} = \int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}\, d\widehat{F}_U,$$

where $\widehat{F}_U$ is the empirical cumulative distribution function obtained from $(U_i)_{i=1,\dots,n_2}$. Let $N = \min(n_1, n_2)$ be the minimum sample size. We assume that, as $n_1, n_2 \to \infty$,

[2] Letting $\lceil a \rceil$ be the least integer greater than or equal to a, notice that, for all $x \in (0,1)$, $\mathbb{G}_{n_1}^{-1}(x) = \xi_{(\lceil n x \rceil)} = F_Y(Y_{(\lceil n x \rceil)})$. By, e.g., Proposition 1.1.3 of Shorack and Wellner (1986), we have almost surely $F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}(x) = F_Y^{-1} \circ F_Y(Y_{(\lceil n x \rceil)}) = Y_{(\lceil n x \rceil)} = \widehat{F}_Y^{-1}(x)$.

$N/n_k \xrightarrow{p} \lambda_k$, $0 \leq \lambda_k \leq 1$, for $k \in \{1, 2\}$. We show that

$$\sqrt{N}(\widehat{\theta}_{n_1,n_2} - \theta_0) = \sqrt{\frac{\lambda_2}{n_2}} \sum_{i=1}^{n_2} \varepsilon_i + \sqrt{\frac{\lambda_1}{n_1}} \sum_{i=1}^{n_1} \eta_i + o_p(1),$$

where $\varepsilon_i = -\int_0^1 [\mathbb{1}\{U_i \leq t\} - F_U(t)]\, dF_Y^{-1}(t)$ and $\eta_i = -\int_0^1 [\mathbb{1}\{F_Y(Y_i) \leq t\} - t] f_U(t)\, dF_Y^{-1}(t)$ are independent, square-integrable, random variables, allowing to apply a standard CLT.

Assumption 5.2.2 (Regularity conditions on densities)

(i) *F_U is absolutely continuous with respect to the Lebesgue measure with a density supported on $[0, 1]$. Moreover, there exist $b_1, b_2 > 0$ and $C_U > 0$ such that for all $t \in (0, 1)$:*

$$f_U(t) \leq C_U t^{-b_1}(1 - t)^{-b_2}.$$

(ii) *There exist $d_1, d_2 > 0$ and $C_Y > 0$ such that for all $t \in (0, 1)$:*

$$|F_Y^{-1}(t)| \leq C_Y t^{-d_1}(1 - t)^{-d_2}.$$

(iii) *$b_1 + d_1 < 1/2$ and $b_2 + d_2 < 1/2$.*

Point 2 of Assumption 5.2.2 holds under the following moment condition on Y.

Lemma 5.2.1 (Lower-level conditions on Y) *Assume $\mathbb{E}[|Y|^p] < \infty$ for $p > 1$, then Assumption 5.2.2(ii) is verified with $d_1 = d_2 = 1/p$.*

Theorem 5.2.2 (Asymptotic normality) *Under Assumptions 5.2.1 and 5.2.2, as $N \to \infty$,*

$$\sqrt{N}(\widehat{\theta}_{n_1,n_2} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

with

$$\sigma^2 = \int_0^1 \int_0^1 \left[\lambda_2 \left[F_U(s \wedge t) - F_U(s)F_U(t)\right] + \lambda_1 [s \wedge t - st] f_U(s) f_U(t)\right] dF_Y^{-1}(s)\, dF_Y^{-1}(t).$$

5.3 Asymptotic Results (Estimated Rank)

In many applications, the random variable U_i has a known form $U_i = F_Z(X_i)$ for some observed random variable X_i and a unknown cumulative distribution function F_Z. As a consequence, we do not directly observe the random variables $(U_i)_{i=1,\dots,n_2}$, instead we are left with estimated quantities $(\widehat{U}_i)_{i=1,\dots,n_2}$. In these cases, $\widehat{U}_i$ is the image of some observed random variable X_i through an empirical cumulative distribution function that comes from another independent sample $(Z_i)_{i=1,\dots,n_3}$, *i.e.* $\widehat{U}_i = \widehat{F}_Z(X_i)$.

Assumption 5.3.1 (Pooled independent samples) $(Y_i)_{i=1,\dots,n_1}$ *(resp.* $(X_i)_{i=1,\dots,n_2}$ *and* $(Z_i)_{i=1,\dots,n_3}$*) are independent draws from the distribution* F_Y

(resp. F_X and F_Z). Moreover, (Y_i, X_j, Z_k) are mutually independent for any value of i,j and k.

Under Assumption 5.2.2(i), U is distributed with cdf $F_U = F_X \circ F_Z^{-1}$ and density

$$f_U(t) = \frac{f_X(F_Z^{-1}(t))}{f_Z(F_Z^{-1}(t))} \mathbb{1}\{t \in [0,1]\}.$$

Notice that Assumption 5.2.2(i) also implies that $\text{Supp}(X) \subseteq \text{Supp}(Z)$ and that F_Z, F_X are continuous. Let $\zeta_j := F_Z(Z_j) \sim \mathcal{U}([0,1])$. Notice that $\widehat{U}_i = \mathbb{H}_{n_3}(U_i)$, where $\mathbb{H}_{n_3}$ is the empirical cdf obtained from $(\zeta_j)_{j=1,\dots,n_3}$. We consider the estimator:

$$\check{\theta}_{n_1,n_2,n_3} := \frac{1}{n_2}\sum_{j=1}^{n_2} \widehat{F}_Y^{-1}(\widehat{U}_j) = \frac{1}{n_2}\sum_{j=1}^{n_2} \widehat{F}_Y^{-1}(\mathbb{H}_{n_3}(U_j)).$$

Remark: *we can also use a smooothed version of $\widehat{F}_Z$. Following Shorack and Wellner (1986), we let $\widehat{F}_Z(Z_{(i)}) = i/(n_3+1)$ for $i = 1, \dots, n_3$, $\widehat{F}_Z(\cdot)$ linear between $Z_{(i)}$ and $Z_{(i+1)}$ for $i < n_3$. For $z < Z_{(1)}$ and $z > Z_{(n_3)}$, we extrapolate linearly until reaching 0 and 1 respectively. One can show that this extrapolation is equivalent to defining $Z_{(0)} = 2Z_{(1)} - Z_{(2)}$ and $Z_{(n_3+1)} = 2Z_{(n_3)} - Z_{(n_3-1)}$ instead of 0 and 1 as in Shorack and Wellner (1986). With this estimator, $\mathbb{H}_{n_3}(\cdot)$ is defined as*

$$\mathbb{H}_{n_3}(u) = \frac{1}{n_3+1}\left(i + \frac{F_Z^{-1}(u) - Z_{(i)}}{Z_{(i+1)} - Z_{(i)}}\right) \quad \text{if } Z_{(i)} \le F_Z^{-1}(u) \le Z_{(i+1)}$$

for $i = 0, \dots, n$. Finally, $\mathbb{H}_{n_3}$ is constant and equal to 0 on $[0, F_Z(Z_{(0)})]$ and constant, equal to 1 on $[F_Z(Z_{(n_3+1)}), 1]$. (these two sets may or may not be empty).

Similarly as before, $\check{\theta}_{n_1,n_2n_3}$ can be expressed as:

$$\check{\theta}_{n_1,n_2,n_3} = \int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U.$$

Here again, we let $N = \min(n_1, n_2, n_3)$ be the minimum sample size, and assume that, as $n_1, n_2, n_3 \to \infty$, $N/n_k \xrightarrow{\text{a.s.}} \lambda_k$, $0 \le \lambda_k \le 1$ for $k \in \{1,2,3\}$ with $\lambda_1 > 0, \lambda_3 > 0$.

Theorem 5.3.1 (Asymptotic normality) *Under Assumptions 5.2.2 and 5.3.1, as $N \to \infty$,*

$$\sqrt{N}(\check{\theta}_{n_1,n_2,n_3} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

with

$$\begin{aligned}\sigma^2 =& \lambda_2 \int_0^1 \int_0^1 \left[F_X(F_Z^{-1}(s \wedge t)) - F_X(F_Z^{-1}(s))F_X(F_Z^{-1}(t))\right] dF_Y^{-1}(s)\, dF_Y^{-1}(t) \\ &+ [\lambda_1 + \lambda_3] \int_0^1 \int_0^1 [s \wedge t - st]\, f_U(s) f_U(t)\, dF_Y^{-1}(s)\, dF_Y^{-1}(t).\end{aligned}$$

The proof is long and technical. Below are the main steps and ingredients:

1. Decompose into three terms. Two are the same as in Theorem 1, the third is new. We decompose it further into several terms: remainder terms plus a L-statistic.

2. For some remainder term, similar technique as in Theorem 1 but a bit more complex. Use in particular the fact that (i) order statistic of uniforms and uniform spacings are distributed as beta; (ii) mean absolute deviation of beta distributions.

3. For another remainder term, use convergence of the supremum of the weighted empirical quantile process (see in particular Csorgo et al., 1986, Corollary 4.3.1).

4. For the $L-$statistic, results in Shorack and Wellner (1986) do not apply here. Instead, we use the necessary and sufficient condition for its asymptotic normality in Hecker (1976).

5.4 Application to Change-in-Change

We study the Change-in-Change estimator of Athey and Imbens (2006). Let $Y_{gt,i}$ the outcome at time t for individual i in group g. The Change-in-Change estimand of the Average Treatment Effect (ATE) is

$$\tau^{CIC} = \mathbb{E}[Y_{11}] - \mathbb{E}[F^{-1}_{Y_{01}}(F_{Y_{00}}(Y_{10}))].$$

The idea is to estimate the counterfactual by averaging the quantile the treated population would have had, had they been in the untreated group at the initial date and kept the same rank in the second period. In our more simplistic framework, we have $U = F_{Y,00}(Y_{10})$, with $Y_{10} \sim F_{Y,10}$. Assume that all the cdf are absolutely continuous with respect to the Lebesgue measure, then $F_U = F_{Y,10} \circ F^{-1}_{Y,00}$ and its density is:

$$f_U(t) = \frac{f_{Y,10}(F^{-1}_{Y,00}(t))}{f_{Y,00}(F^{-1}_{Y,00}(t))} \mathbb{1}\{t \in [0,1]\}.$$

Clearly, if the outcome distribution is the same for the treated and the untreated at the initial date ($f_{Y,10} = f_{Y,00}$) then U is uniformly distributed. Athey and Imbens (2006) require the density of Y_{gt} for each g and t to be bounded from below and bounded from above on a compact support (see Assumption 5 therein). This assumption yields that f_U will also be bounded. In general, f_U will be bounded if and only if the ratio $f_{Y,10}/f_{Y,00}$ is bounded, which may not be the case for many usual distributions typically encountered with economic data. Our method of proof does not require any constant bound on f_U, thus extending the cases where the Change-in-Change is a relevant tool.

Examples: In the following examples we study the tail behavior of f_U with respect to the underlying distribution of the treated and untreated outcomes at the initial date. We show that under many standard distributions, Assumption 5.2.2(i) is verified.

1. Exponential Distribution. Assume that $Y_{g0} \sim \mathcal{E}(\lambda_g)$, in that case

$$f_U(t) = \frac{\lambda_1}{\lambda_0}(1-t)^{\lambda_1/\lambda_0 - 1}\mathbb{1}\{t \in [0,1]\},$$

and $U \sim \text{Beta}(1, \lambda_1/\lambda_0)$.

2. Pareto Distribution. Assume that Y_{g0} has cdf $1-(\beta_g/x)^{\alpha_g}$, in that case

$$f_U(t) = \frac{\alpha_1}{\alpha_0}\left(\frac{\beta_1}{\beta_0}\right)^{\alpha_1}(1-t)^{\alpha_1/\alpha_0 - 1}\mathbb{1}\{1-(\beta_0/\beta_1)^{\alpha_0} < t < 1\},$$

which is a "truncated" Beta distribution.

3. Normal Distribution. Assume that $Y_{g0} \sim \mathcal{N}(\mu_g, \sigma_g^2)$, in that case

$$f_U(t) = \frac{\sigma_0}{\sigma_1}\exp\left[-\frac{1}{2\sigma_1^2}\left((\sigma_0+\sigma_1)\Phi^{-1}(t)+\mu_0-\mu_1\right)\left((\sigma_0-\sigma_1)\Phi^{-1}(t)+\mu_0-\mu_1\right)\right] \times \mathbb{1}\{t \in [0,1]\}.$$

If $\sigma_1 < \sigma_0$, $f_U(t) \to 0$, when either $t \to 0$ or $t \to 1$. Now if $\sigma_1 > \sigma_0$, the analysis is more complicated. Consider the special case: $\mu_1 = \mu_0$. For $t \in (1/2, 1)$, using the inequality $\Phi^{-1}(t) \leq \sqrt{-2\ln(2(1-t))}$ yields $f_U(t) \leq (\sigma_0/\sigma_1)(2(1-t))^{\sigma_0^2/\sigma_1^2 - 1}$. Symmetrically, for $t \in (0, 1/2)$, $\Phi^{-1}(t) \geq -\sqrt{-2\ln(2t)}$ yields $f_U(t) \leq (\sigma_0/\sigma_1)(2t)^{\sigma_0^2/\sigma_1^2-1}$.

4. Logistic Distribution. Assume that Y_{g0} has cdf $1/(1+\exp(-(t-\mu_g)/\beta_g))$, in that case

$$f_U(t) = \frac{\beta_0}{\beta_1}\frac{(1/t-1)^{\beta_0/\beta_1 - 1}\, e^{(\mu_1-\mu_0)/\beta_1}}{t^2\left(1+(1/t-1)^{\beta_0/\beta_1}\, e^{(\mu_1-\mu_0)/\beta_1}\right)^2}\mathbb{1}\{t \in [0,1]\}.$$

5. Gumbel Distribution. Assume that Y_{g0} has cdf $e^{-(t-\mu_g)/\beta_g}\exp\left(-e^{-(t-\mu_g)/\beta_g}\right)/\beta_g$, in that case

$$f_U(t) = \frac{\beta_0}{\beta_1}e^{(\mu_1-\mu_0)/\beta_1}\ln(t)^{\beta_0/\beta_1 - 1}\exp\left(-\ln(t)^{\beta_0/\beta_1}e^{(\mu_1-\mu_0)/\beta_1}\right)\mathbb{1}\{t \in [0,1]\}.$$

If $\beta_1 = \beta_0 = 1$ and $\mu_0 > \mu_1$, $U \sim \text{Beta}(1-e^{\mu_1-\mu_0}, 1)$.

We never observe U_i directly, instead F_{00} is replaced by its empirical counterpart $\widehat{F}_{00}$ and we have $\widehat{U}_i = \widehat{F}_{00}(Y_{i,10})$. Notice that:

$$\begin{aligned}\widehat{U}_i &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\{Y_{00,j} \leq Y_{10,i}\} \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}\{F_{00}(Y_{00,j}) \leq U_i\} \\ &= \mathbb{H}_{n_3}(U_i),\end{aligned}$$

where $\mathbb{H}_{n_3}$ is the empirical cdf of $F_{00}(Y_{00,j}) \sim \mathcal{U}[0,1]$.

Theorem 2 generalizes Theorem 5.1 in Athey and Imbens (2006). Recall the definition of η_i, φ_i, and ε_i in the proof of Theorem 5.3.1. By several changes of variables, one may easily verify that

$$\eta_i = -\int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} \left[\mathbb{1}\{F_Y(Y_i) \leq F_Z(x)\} - F_Z(x)\right] \frac{1}{f_Y\left(F_Y^{-1}(F_Z(x))\right)} \times f_X(x)\,dx, \tag{5.4.1}$$

$$\varphi_i = \int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} \left[\mathbb{1}\{Z_i \leq x\} - F_Z(x)\right] \frac{1}{f_Y\left(F_Y^{-1}(F_Z(x))\right)} \times f_X(x)\,dx. \tag{5.4.2}$$

Under Assumption 5.2.2, we have $\mathbb{E}(\zeta_i^2) = V^q$ and $\mathbb{E}(\varphi_i^2) = V^p$, where V^q and V^p are defined in Theorem 5.1 in Athey and Imbens (2006). It thus remains to analyze the last term, V^r, appearing in the asymptotic variance of their Theorem. We have:

$$\begin{aligned} V^r &:= \mathbb{V}\left(F_Y^{-1}\left(F_Z(X)\right)\right) \\ &= \int_{Supp(X)} \left[F_Y^{-1}\left(F_Z\left(x\right)\right)\right]^2 f_X(x)\,dx - \theta_0^2. \end{aligned}$$

By an integration by part

$$\begin{aligned} \mathbb{E}(\varepsilon_i^2) &= \int_0^1 \int_0^1 \left[F_X(F_Z^{-1}(s \wedge t)) - F_X(F_Z^{-1}(s))F_X(F_Z^{-1}(t))\right] dF_Y^{-1}(s)\, dF_Y^{-1}(t) \\ &= \int_0^1 \Big\{ \left[F_Y^{-1}(s)(F_X(F_Z^{-1}(s \wedge t)) - F_X(F_Z^{-1}(s))F_X(F_Z^{-1}(t)))\right]_{s=0}^{s=1} \\ &\quad - \int_0^1 F_Y^{-1}(s)\Big[\frac{f_X(F_Z^{-1}(s \wedge t))}{f_Z(F_Z^{-1}(s \wedge t))}\mathbb{1}\{s \leq t\} - \frac{f_X(F_Z^{-1}(s))F_X(F_Z^{-1}(t))}{f_Z(F_Z^{-1}(s))}\Big]\, ds \Big\}\, dF_Y^{-1}(t) \\ &= -\int_0^1 \left(\int_0^1 F_Y^{-1}(s) \frac{f_X(F_Z^{-1}(s \wedge t))}{f_Z(F_Z^{-1}(s \wedge t))} \mathbb{1}\{s \leq t\}\, ds \right) dF_Y^{-1}(t) \\ &\quad + \int_0^1 \left(\int_0^1 F_Y^{-1}(s) \frac{f_X(F_Z^{-1}(s))F_X(F_Z^{-1}(t))}{f_Z(F_Z^{-1}(s))}\, ds \right) dF_Y^{-1}(t). \end{aligned} \tag{5.4.3}$$

The third equality follows because Assumption 5.2.2 implies

$$\left[F_Y^{-1}(s)(F_X(F_Z^{-1}(s \wedge t)) - F_X(F_Z^{-1}(s))F_X(F_Z^{-1}(t)))\right]_{s=0}^{s=1} = 0.$$

Focus on the second term in (5.4.3). By the change of variable $x = F_Z^{-1}(s)$, we obtain

$$\begin{aligned}
&\int_0^1 \left(\int_0^1 F_Y^{-1}(s) \frac{f_X(F_Z^{-1}(s)) F_X(F_Z^{-1}(t))}{f_Z(F_Z^{-1}(s))}\, ds \right) dF_Y^{-1}(t) \\
&= \left(\int_0^1 F_Y^{-1}(s) \frac{f_X(F_Z^{-1}(s))}{f_Z(F_Z^{-1}(s))}\, ds \right) \left(\int_0^1 F_X(F_Z^{-1}(t))\, dF_Y^{-1}(t) \right) \\
&= \left(\int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right) \left(\int_0^1 F_X(F_Z^{-1}(t))\, dF_Y^{-1}(t) \right) \\
&= \left(\int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right) \left([F_Y^{-1}(t) F_X(F_Z^{-1}(t))]_{t=0}^{t=1} \right. \\
&\qquad \left. - \int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right) \\
&= - \left(\int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right)^2 + \left(\int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right) F_Y^{-1}(1) \\
&= -\theta_0^2 + \theta_0 F_Y^{-1}(1), \qquad (5.4.4)
\end{aligned}$$

where we obtain the third equality by an integration by part followed by the same change of variable than before, and the last two equalities hold by Assumption 5.2.2. Now, focus on the first term in (5.4.3). By the same change of variable again, we have

$$\begin{aligned}
&- \int_0^1 \left(\int_0^1 F_Y^{-1}(s) \frac{f_X(F_Z^{-1}(s \wedge t))}{f_Z(F_Z^{-1}(s \wedge t))} \mathbb{1}\{s \le t\}\, ds \right) dF_Y^{-1}(t) \\
&\quad = - \int_0^1 \left(\int_{F_Z^{-1}(0)}^{F_Z^{-1}(t)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right) dF_Y^{-1}(t). \qquad (5.4.5)
\end{aligned}$$

By Leibniz's derivation rule for integrals, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{dt}} \left(\int_{F_Z^{-1}(0)}^{F_Z^{-1}(t)} F_Y^{-1}(F_Z(x)) f_X(x)\, dx \right) &= \frac{1}{f_Z(F_Z^{-1}(t))} F_Y^{-1}(F_Z(F_Z^{-1}(t))) f_X(F_Z^{-1}(t)) \\
&= \frac{f_X(F_Z^{-1}(t))}{f_Z(F_Z^{-1}(t))} F_Y^{-1}(t).
\end{aligned}$$

Hence, an integration by part of (5.4.5) yields

$$\begin{aligned}
&- \int_0^1 \left(\int_0^1 F_Y^{-1}(s) \frac{f_X(F_Z^{-1}(s \wedge t))}{f_Z(F_Z^{-1}(s \wedge t))} \mathbb{1}\{s \le t\}\, ds \right) dF_Y^{-1}(t) \\
&\quad = -\theta_0 F_Y^{-1}(1) + \int_0^1 F_Y^{-1}(t)^2 \frac{f_X(F_Z^{-1}(t))}{f_Z(F_Z^{-1}(t))}\, dt \\
&\quad = -\theta_0 F_Y^{-1}(1) + \int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} \left[F_Y^{-1}(F_Z(x)) \right]^2 f_X(x)\, dx, \qquad (5.4.6)
\end{aligned}$$

where we used the change of variable $x = F_Z^{-1}(t)$ to obtain the last equality. Now, by combining (5.4.4) with (5.4.6), and noting that Assumption 5.2.2 implies

$$\int_{F_Z^{-1}(0)}^{F_Z^{-1}(1)} \left[F_Y^{-1}(F_Z(x))\right]^2 f_X(x)\,dx = \int_{Supp(X)} \left[F_Y^{-1}(F_Z(x))\right]^2 f_X(x)\,dx,$$

we obtain

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= -F_Y^{-1}(1)\theta_0 + \int_{Supp(X)} \left[F_Y^{-1}(F_Z(x))\right]^2 f_X(x)\,dx - \theta_0^2 + \theta_0 F_Y^{-1}(1) \\ &= \int_{Supp(X)} \left[F_Y^{-1}(F_Z(x))\right]^2 f_X(x)\,dx - \theta_0^2 \\ &= V^r.\end{aligned}$$

5.5 Monte Carlo Simulations

In the following experiments, we consider two types of estimators for θ_0 that differ in their way of computing $\widehat{U}_i = \widehat{F}_Z(X_i)$:

1. The first one considers a standard estimator for the cdf of Z : $\widehat{F}_Z(t) = n_Z^{-1} \sum_{i=1}^{n_Z} \mathbb{1}\{t \leq Z_i\}$. We call the resulting estimator *standard*.

2. The second one considers a piece-wise linear version of the cdf estimator, where a local linear approximation is computed between two consecutive points, as in Shorack and Wellner (1986). For example, for a point t in the interval $(Z_{(i)}, Z_{(i+1)}]$, instead of a constant value of i/n_Z, the function evaluates to :

$$\frac{1}{n_Z + 1}\left[i + \frac{t - Z_{(i)}}{Z_{(i+1)} - Z_{(i)}}\right].$$

   We call the resulting estimator *smoothed*.

Similarly, several options are considered for estimating the standard error. They differ in their way of estimating the function $t \to 1/f_Y\left(F_Y^{-1}(F_Z(t))\right)$ that appears in the expressions for η_i and φ_i defined in Theorem 5.3.1 :

1. The first considers a version where f_Y is estimated using a kernel estimator, as in Athey and Imbens (2006). Here we opt for a Gaussian kernel. The function F_Y^{-1} is estimated using a standard quantile function, and F_Z is estimated either by the standard or the smoothed cdf, in congruence with the estimator above. This version of the standard error estimator is called *kernel*.

2. The second version follows an idea developed in Lewbel and Schennach (2007) and relies on observing that $1/f_Y\left(F_Y^{-1}\right)$ is the derivative of F_Y^{-1} and as such can be estimated by $(F_Y^{-1}(t_{(i+1)}) - F_Y^{-1}(t_{(i)}))/(t_{(i+1)} - t_{(i)})$ for two consecutive data points $t_{(i)}$ and $t_{(i+1)}$. In practice, the ranks $\widehat{U}_i = \widehat{F}_Z(X_i)$ are first computed. Duplicated values are then discarded to make sure that once reordered,

consecutive values are not equal. Then, the function $t \to 1/f_Y\left(F_Y^{-1}(F_Z(t))\right)$ evaluated at $X_{(i)}$ is estimated by $(F_Y^{-1}(\hat{U}_{(i+1)}) - F_Y^{-1}(\hat{U}_{(i)}))/(\hat{U}_{(i+1)} - \hat{U}_{(i)})$. This version of the standard error estimator is called *Lewbel-Schennach*.

3. The third version draws from the same source for inspiration, without deleting duplicated values of $\widehat{U}_i$. Instead, $t \to 1/f_Y\left(F_Y^{-1}(F_Z(t))\right)$ evaluated at $X_{(i)}$ is estimated by $(F_Y^{-1}(\hat{U}_{[i+1]}) - F_Y^{-1}(\hat{U}_{[i-1]}))/(\hat{U}_{[i+1]} - \hat{U}_{[i-1]})$, where $\hat{U}_{[i+1]}$ (resp. $\hat{U}_{[i-1]}$) indicates the smallest (resp. largest) value inside the sample $(\hat{U}_i)_i$ strictly greater (resp. smaller) than $\hat{U}_i$. Extreme values are handled by taking the values themselves when a larger / smaller value cannot be found in the sample. This version of the standard error estimator is called *Xavier*.

5.5.1 Exponential-Pareto DGP

The first experiment favors computational tractability to tightly control cases where Assumption 5.2.2 is violated, with the hope of finding evidence that it is a necessary condition for Theorem 5.3.1 to hold. For that purpose, it assumes that both X and Z have exponential distributions : $X \sim \mathcal{E}(\lambda_X)$ and $Z \sim \mathcal{E}(1)$. As a consequence, U has density $\lambda_X(1-t)^{\lambda_X - 1}\mathbb{1}\{t \in [0,1]\}$. Y has a Pareto distribution re-scaled to the real line, *i.e.* density $\alpha_Y(1+t)^{-(\alpha_Y+1)}\mathbb{1}\{t > 0\}$ and quantile function $(1-t)^{-1/\alpha_Y} - 1$. The corresponding parameters in Assumption 5.2.2 (i) become $b_1 = 0$ and $b_2 = 1 - \lambda_X$ while those in Assumption 5.2.2 (ii) become $d_1 = 0$ and $d_2 = 1/\alpha_Y$. As a consequence, the constraint $b_1 + d_1 < 1/2$ is slacking while $b_2 + d_2 < 1/2$ can be violated depending on the parameter values, meaning that the issue lies at the right tail of the distributions. Notice that when $b_2 + d_2 > 1$, the parameter of interest is not defined so we exclude those cases by considering DGPs such that $b_2 + d_2 \in (0,1)$. Theorem 5.3.1 implies that when $b_2 + d_2 < 1/2$, the estimator is asymptotically Gaussian. If Assumption 5.2.2 is necessary, cases where $b_2 + d_2 \geq 1/2$ should result in a non-gaussian behavior for the same estimator.

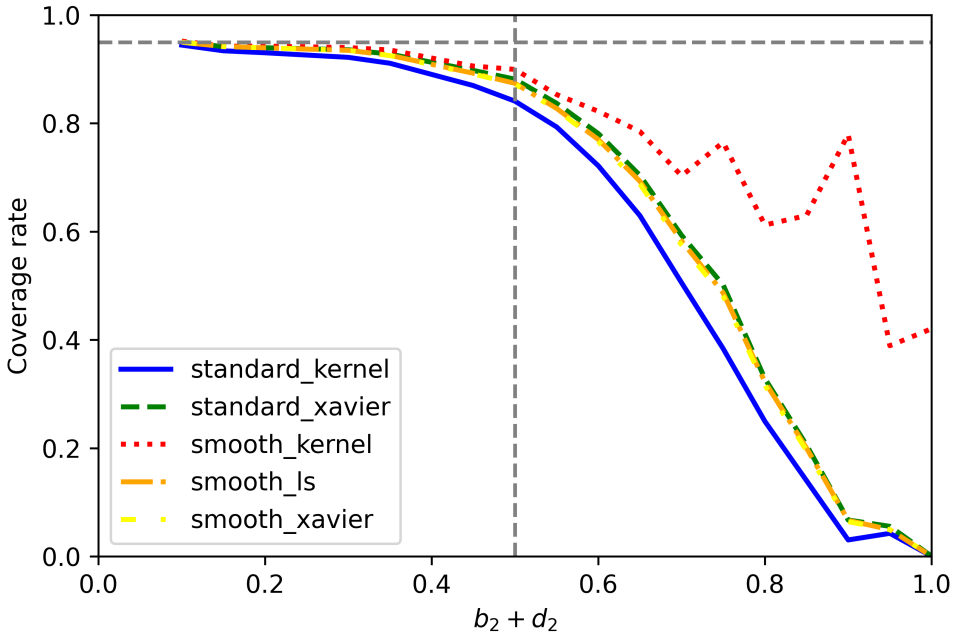
We run $10,000$ simulations for three different sample sizes : 100, 500 and 1000. Values of $\lambda_X = 1 - b_2$ across DGPs span the set $\{.05, .1, .25, .4, .5, .6, .75, .9, .95\}$, while values of $\alpha_Y = 1/d_2$ span the set $\{20, 10, 4, 2.5, 2, 1.666, 1.333, 1.111, 1.053\}$, so that both b_2 and d_2 span the same set. DGPs where $b_2 + d_2 > 1$ are discarded. In the end, this experiment spans 37 different values of the couple (λ_X, α_Y) for three sample sizes.

Figure 5.1 plots the coverage rates as a function of the sum $b_2 + d_2$.

5.5.2 Gaussian DGP

In this experiment, both Z and Y are distributed as $\mathcal{N}(0,1)$, while X is distributed as $\mathcal{N}(\mu_X, \sigma_X^2)$, which implies $\theta_0 = \mu_X$, so the parameter of interest is always defined. If $\sigma_X < 1$, $f_U(t) \to 0$, when either $t \to 0$ or $t \to 1$, so this case does not raise any concern. However, when $\sigma_X > 1$, the analysis is more complicated. In the special case where $\mu_X = 0$. For $t \in (1/2, 1)$, using the inequality $\Phi^{-1}(t) \leq \sqrt{-2\ln(2(1-t))}$

FIGURE 5.1: Exponential DGP, Coverage Rates as a Function of $b_2 + d_2$ – Sample Size = 1000.



yields $f_U(t) \leq (1/\sigma_X)(2(1-t))^{1/\sigma_X^2-1}$. Symmetrically, for $t \in (0, 1/2)$, $\Phi^{-1}(t) \geq -\sqrt{-2\ln(2t)}$ yields $f_U(t) \leq (1/\sigma_X)(2t)^{1/\sigma_X^2-1}$. So this DGP implies $b_1 = b_2 = 1 - 1/\sigma_X^2$. And from Lemma 5.2.1, we can deduce $d_1 = d_2 \approx 0$. As a consequence, σ_X^2 is the key parameter that should govern the asymptotic behavior of the estimator. The intuition is that Theorem 5.3.1 should apply whenever $\sigma_X^2 < 2$.

We run $10,000$ simulations for three different sample sizes : 100, 500 and 1000. Values of $\sigma_X^2 = 1/(1 - b_2)$ across DGPs span the set $\{20, 10, 4, 2.5, 2, 1.666, 1.333, 1.111, 1.053\}$, while values of μ_X span the set $\{0, .5, 1, 2, 3\}$. In the end, this experiment spans 45 different values of the couple (μ_X, σ_X^2) for three sample sizes.

TABLE 5.1: Gaussian simulations, $B = 10,000$.

	n=100				n=500				n=1000			
Estimator	Bias	RMSE	Cov. rate	CI size	Bias	RMSE	Cov. rate	CI size	Bias	RMSE	Cov. rate	CI size
					μ_X=0, σ_X^2=20 – $b_1 = b_2$=0.95							
standard kernel	-0.0	0.297	0.853	0.852	0.0	0.186	0.786	0.45	-0.0	0.154	0.733	0.337
standard xavier	-0.0	0.297	0.855	0.858	0.0	0.186	0.794	0.458	-0.0	0.154	0.745	0.345
smooth kernel	-0.0	0.293	0.86	0.853	0.0	0.183	0.797	0.455	-0.0	0.152	0.749	0.348
smooth ls	-0.0	0.293	0.917	1.009	0.0	0.183	0.863	0.537	-0.0	0.152	0.82	0.408
smooth xavier	-0.0	0.293	0.863	0.862	0.0	0.183	0.806	0.463	-0.0	0.152	0.756	0.35
					μ_X=0, σ_X^2=10 – $b_1 = b_2$=0.9							
standard kernel	0.003	0.267	0.878	0.815	0.0	0.154	0.835	0.424	-0.001	0.122	0.808	0.317
standard xavier	0.003	0.267	0.882	0.825	0.0	0.154	0.847	0.435	-0.001	0.122	0.824	0.328
smooth kernel	0.003	0.262	0.886	0.817	0.0	0.151	0.847	0.429	-0.001	0.12	0.824	0.327
smooth ls	0.003	0.262	0.929	0.945	0.0	0.151	0.896	0.491	-0.001	0.12	0.88	0.371
smooth xavier	0.003	0.262	0.89	0.83	0.0	0.151	0.858	0.441	-0.001	0.12	0.837	0.334
					μ_X=0, σ_X^2=4 – $b_1 = b_2$=0.75							
standard kernel	0.002	0.221	0.916	0.756	-0.0	0.112	0.908	0.377	-0.0	0.083	0.908	0.277
standard xavier	0.002	0.221	0.921	0.769	-0.0	0.112	0.918	0.389	-0.0	0.083	0.918	0.287
smooth kernel	0.002	0.217	0.922	0.757	-0.0	0.11	0.916	0.381	-0.0	0.082	0.915	0.281
smooth ls	0.002	0.217	0.947	0.833	-0.0	0.11	0.937	0.411	-0.0	0.082	0.936	0.301
smooth xavier	0.002	0.217	0.927	0.772	-0.0	0.11	0.925	0.391	-0.0	0.082	0.925	0.29
					μ_X=0, σ_X^2=2.5 – $b_1 = b_2$=0.6							
standard kernel	-0.001	0.2	0.93	0.724	-0.001	0.096	0.932	0.349	-0.001	0.069	0.931	0.253
standard xavier	-0.001	0.2	0.934	0.734	-0.001	0.096	0.936	0.357	-0.001	0.069	0.937	0.259
smooth kernel	-0.001	0.196	0.935	0.723	-0.001	0.095	0.935	0.349	-0.001	0.068	0.934	0.255
smooth ls	-0.001	0.196	0.946	0.768	-0.001	0.095	0.946	0.366	-0.001	0.068	0.943	0.264
smooth xavier	-0.001	0.196	0.939	0.734	-0.001	0.095	0.94	0.358	-0.001	0.068	0.938	0.26
					μ_X=0, σ_X^2=2 – $b_1 = b_2$=0.5							
standard kernel	0.001	0.192	0.936	0.709	0.001	0.091	0.936	0.336	0.001	0.064	0.942	0.241
standard xavier	0.001	0.192	0.937	0.716	0.001	0.091	0.94	0.342	0.001	0.064	0.946	0.246
smooth kernel	0.001	0.188	0.941	0.708	0.001	0.09	0.939	0.336	0.001	0.063	0.944	0.241
smooth ls	0.001	0.188	0.948	0.739	0.001	0.09	0.945	0.347	0.001	0.063	0.951	0.248
smooth xavier	0.001	0.188	0.941	0.715	0.001	0.09	0.942	0.342	0.001	0.063	0.949	0.246

	μ_X=0, σ^2_X=1.666 – $b_1 = b_2$=0.4											
standard kernel	0.0	0.185	0.939	0.699	-0.0	0.086	0.943	0.325	0.0	0.061	0.942	0.233
standard xavier	0.0	0.185	0.94	0.702	-0.0	0.086	0.946	0.33	0.0	0.061	0.946	0.236
smooth kernel	0.0	0.181	0.943	0.697	-0.0	0.085	0.945	0.325	0.0	0.06	0.944	0.235
smooth ls	0.0	0.181	0.95	0.718	-0.0	0.085	0.949	0.332	0.0	0.06	0.949	0.237
smooth xavier	0.0	0.181	0.945	0.701	-0.0	0.085	0.947	0.329	0.0	0.06	0.947	0.236
	μ_X=0, σ^2_X=1.333 – $b_1 = b_2$=0.25											
standard kernel	-0.002	0.18	0.942	0.685	0.001	0.081	0.946	0.314	-0.0	0.058	0.945	0.224
standard xavier	-0.002	0.18	0.942	0.683	0.001	0.081	0.946	0.316	-0.0	0.058	0.946	0.225
smooth kernel	-0.002	0.176	0.946	0.683	0.001	0.081	0.946	0.314	-0.0	0.058	0.946	0.223
smooth ls	-0.002	0.176	0.948	0.692	0.001	0.081	0.949	0.317	-0.0	0.058	0.947	0.225
smooth xavier	-0.002	0.176	0.944	0.681	0.001	0.081	0.948	0.316	-0.0	0.058	0.946	0.225
	μ_X=0, σ^2_X=1.111 – $b_1 = b_2$=0.1											
standard kernel	-0.001	0.175	0.948	0.676	-0.001	0.079	0.949	0.307	0.001	0.056	0.947	0.217
standard xavier	-0.001	0.175	0.945	0.671	-0.001	0.079	0.949	0.307	0.001	0.056	0.947	0.218
smooth kernel	-0.0	0.172	0.952	0.674	-0.001	0.078	0.95	0.306	0.001	0.056	0.948	0.217
smooth ls	-0.0	0.172	0.95	0.676	-0.001	0.078	0.951	0.307	0.001	0.056	0.948	0.218
smooth xavier	-0.0	0.172	0.946	0.668	-0.001	0.078	0.951	0.306	0.001	0.056	0.948	0.217
	μ_X=0, σ^2_X=1.053 – $b_1 = b_2$=0.05											
standard kernel	0.001	0.174	0.944	0.673	-0.0	0.077	0.95	0.305	0.001	0.056	0.946	0.216
standard xavier	0.001	0.174	0.942	0.666	-0.0	0.077	0.95	0.304	0.001	0.056	0.946	0.216
smooth kernel	0.001	0.171	0.948	0.671	-0.0	0.077	0.951	0.304	0.001	0.056	0.947	0.216
smooth ls	0.001	0.171	0.947	0.671	-0.0	0.077	0.952	0.304	0.001	0.056	0.948	0.216
smooth xavier	0.001	0.171	0.945	0.664	-0.0	0.077	0.952	0.304	0.001	0.056	0.948	0.215
	μ_X=0.5, σ^2_X=20 – $b_1 = b_2$=0.95											
standard kernel	-0.292	0.295	0.659	0.848	-0.249	0.188	0.424	0.449	-0.237	0.157	0.31	0.336
standard xavier	-0.292	0.295	0.662	0.854	-0.249	0.188	0.431	0.456	-0.237	0.157	0.318	0.344
smooth kernel	-0.293	0.291	0.66	0.85	-0.25	0.185	0.427	0.456	-0.238	0.154	0.314	0.354
smooth ls	-0.293	0.291	0.75	1.009	-0.25	0.185	0.516	0.54	-0.238	0.154	0.396	0.411
smooth xavier	-0.293	0.291	0.665	0.859	-0.25	0.185	0.437	0.463	-0.238	0.154	0.324	0.351
	μ_X=0.5, σ^2_X=10 – $b_1 = b_2$=0.9											
standard kernel	-0.214	0.267	0.746	0.809	-0.172	0.154	0.576	0.421	-0.157	0.124	0.481	0.315
standard xavier	-0.214	0.267	0.75	0.819	-0.172	0.154	0.589	0.432	-0.157	0.124	0.498	0.326
smooth kernel	-0.218	0.262	0.749	0.815	-0.174	0.151	0.581	0.43	-0.159	0.121	0.488	0.327

smooth ls	-0.218	0.262	0.815	0.953	-0.174	0.151	0.66	0.497	-0.159	0.121	0.566	0.376
smooth xavier	-0.218	0.262	0.755	0.83	-0.174	0.151	0.594	0.442	-0.159	0.121	0.506	0.336
						μ_X=0.5, σ_X^2=4 – $b_1 = b_2$=0.75						
standard kernel	-0.112	0.225	0.855	0.748	-0.071	0.117	0.821	0.376	-0.059	0.086	0.807	0.277
standard xavier	-0.112	0.225	0.862	0.762	-0.071	0.117	0.834	0.389	-0.059	0.086	0.823	0.288
smooth kernel	-0.119	0.22	0.859	0.76	-0.076	0.115	0.823	0.385	-0.063	0.085	0.81	0.302
smooth ls	-0.119	0.22	0.897	0.849	-0.076	0.115	0.861	0.424	-0.063	0.085	0.844	0.312
smooth xavier	-0.119	0.22	0.866	0.779	-0.076	0.115	0.838	0.4	-0.063	0.085	0.826	0.298
						μ_X=0.5, σ_X^2=2.5 – $b_1 = b_2$=0.6						
standard kernel	-0.069	0.204	0.899	0.717	-0.036	0.098	0.902	0.349	-0.025	0.072	0.903	0.255
standard xavier	-0.069	0.204	0.905	0.73	-0.036	0.098	0.912	0.36	-0.025	0.072	0.912	0.263
smooth kernel	-0.079	0.199	0.903	0.731	-0.042	0.097	0.904	0.356	-0.03	0.071	0.905	0.264
smooth ls	-0.079	0.199	0.923	0.793	-0.042	0.097	0.921	0.38	-0.03	0.071	0.92	0.276
smooth xavier	-0.079	0.199	0.907	0.749	-0.042	0.097	0.913	0.369	-0.03	0.071	0.914	0.27
						μ_X=0.5, σ_X^2=2 – $b_1 = b_2$=0.5						
standard kernel	-0.05	0.197	0.914	0.703	-0.02	0.094	0.918	0.338	-0.014	0.068	0.924	0.245
standard xavier	-0.05	0.197	0.917	0.715	-0.02	0.094	0.926	0.347	-0.014	0.068	0.931	0.252
smooth kernel	-0.061	0.192	0.918	0.717	-0.026	0.092	0.923	0.345	-0.018	0.067	0.926	0.251
smooth ls	-0.061	0.192	0.933	0.767	-0.026	0.092	0.935	0.361	-0.018	0.067	0.936	0.26
smooth xavier	-0.061	0.192	0.922	0.733	-0.026	0.092	0.93	0.354	-0.018	0.067	0.933	0.257
						μ_X=0.5, σ_X^2=1.666 – $b_1 = b_2$=0.4						
standard kernel	-0.035	0.19	0.928	0.693	-0.015	0.089	0.925	0.329	-0.008	0.064	0.934	0.237
standard xavier	-0.035	0.19	0.93	0.702	-0.015	0.089	0.932	0.337	-0.008	0.064	0.94	0.243
smooth kernel	-0.047	0.186	0.932	0.707	-0.021	0.088	0.928	0.333	-0.012	0.063	0.937	0.242
smooth ls	-0.047	0.186	0.94	0.747	-0.021	0.088	0.938	0.347	-0.012	0.063	0.946	0.249
smooth xavier	-0.047	0.186	0.933	0.719	-0.021	0.088	0.933	0.342	-0.012	0.063	0.944	0.246
						μ_X=0.5, σ_X^2=1.333 – $b_1 = b_2$=0.25						
standard kernel	-0.027	0.186	0.93	0.683	-0.007	0.084	0.94	0.319	-0.005	0.061	0.941	0.228
standard xavier	-0.027	0.186	0.93	0.688	-0.007	0.084	0.944	0.325	-0.005	0.061	0.945	0.233
smooth kernel	-0.039	0.182	0.937	0.696	-0.012	0.083	0.942	0.324	-0.008	0.06	0.944	0.231
smooth ls	-0.039	0.182	0.941	0.723	-0.012	0.083	0.949	0.332	-0.008	0.06	0.949	0.236
smooth xavier	-0.039	0.182	0.935	0.703	-0.012	0.083	0.946	0.329	-0.008	0.06	0.948	0.235
						μ_X=0.5, σ_X^2=1.111 – $b_1 = b_2$=0.1						
standard kernel	-0.016	0.178	0.94	0.675	-0.006	0.082	0.944	0.312	-0.003	0.058	0.946	0.223
standard xavier	-0.016	0.178	0.94	0.677	-0.006	0.082	0.947	0.316	-0.003	0.058	0.949	0.226

smooth kernel	-0.028	0.175	0.945	0.687	-0.01	0.081	0.946	0.315	-0.005	0.058	0.949	0.224
smooth ls	-0.028	0.175	0.948	0.706	-0.01	0.081	0.949	0.321	-0.005	0.058	0.953	0.228
smooth xavier	-0.028	0.175	0.944	0.691	-0.01	0.081	0.948	0.319	-0.005	0.058	0.952	0.227
	μ_X=0.5, σ_X^2=1.053 – $b_1 = b_2$=0.05											
standard kernel	-0.017	0.178	0.938	0.674	-0.004	0.082	0.943	0.311	-0.002	0.058	0.943	0.221
standard xavier	-0.017	0.178	0.937	0.676	-0.004	0.082	0.943	0.314	-0.002	0.058	0.945	0.224
smooth kernel	-0.029	0.174	0.946	0.686	-0.008	0.081	0.944	0.313	-0.005	0.058	0.944	0.222
smooth ls	-0.029	0.174	0.948	0.704	-0.008	0.081	0.946	0.319	-0.005	0.058	0.947	0.226
smooth xavier	-0.029	0.174	0.944	0.69	-0.008	0.081	0.945	0.317	-0.005	0.058	0.947	0.225
	μ_X=1, σ_X^2=20 – $b_1 = b_2$=0.95											
standard kernel	-0.578	0.299	0.281	0.839	-0.502	0.19	0.079	0.444	-0.472	0.16	0.039	0.333
standard xavier	-0.578	0.299	0.284	0.845	-0.502	0.19	0.082	0.451	-0.472	0.16	0.042	0.341
smooth kernel	-0.581	0.295	0.278	0.842	-0.504	0.187	0.08	0.452	-0.474	0.157	0.041	0.36
smooth ls	-0.581	0.295	0.378	1.01	-0.504	0.187	0.122	0.541	-0.474	0.157	0.067	0.414
smooth xavier	-0.581	0.295	0.286	0.852	-0.504	0.187	0.084	0.46	-0.474	0.157	0.044	0.35
	μ_X=1, σ_X^2=10 – $b_1 = b_2$=0.9											
standard kernel	-0.445	0.269	0.405	0.792	-0.35	0.161	0.185	0.413	-0.317	0.131	0.113	0.31
standard xavier	-0.445	0.269	0.412	0.802	-0.35	0.161	0.195	0.425	-0.317	0.131	0.123	0.321
smooth kernel	-0.45	0.264	0.402	0.802	-0.354	0.158	0.186	0.427	-0.321	0.129	0.116	0.791
smooth ls	-0.45	0.264	0.509	0.951	-0.354	0.158	0.26	0.501	-0.321	0.129	0.174	0.382
smooth xavier	-0.45	0.264	0.414	0.818	-0.354	0.158	0.202	0.439	-0.321	0.129	0.129	0.336
	μ_X=1, σ_X^2=4 – $b_1 = b_2$=0.75											
standard kernel	-0.244	0.234	0.669	0.72	-0.153	0.124	0.578	0.369	-0.125	0.097	0.524	0.275
standard xavier	-0.244	0.234	0.678	0.737	-0.153	0.124	0.602	0.385	-0.125	0.097	0.552	0.29
smooth kernel	-0.258	0.228	0.668	0.746	-0.162	0.122	0.576	0.392	-0.133	0.095	0.521	0.298
smooth ls	-0.258	0.228	0.738	0.865	-0.162	0.122	0.648	0.441	-0.133	0.095	0.596	0.331
smooth xavier	-0.258	0.228	0.68	0.772	-0.162	0.122	0.604	0.408	-0.133	0.095	0.553	0.308
	μ_X=1, σ_X^2=2.5 – $b_1 = b_2$=0.6											
standard kernel	-0.154	0.214	0.783	0.689	-0.081	0.11	0.774	0.349	-0.062	0.083	0.768	0.259
standard xavier	-0.154	0.214	0.791	0.708	-0.081	0.11	0.794	0.367	-0.062	0.083	0.789	0.274
smooth kernel	-0.173	0.21	0.784	0.722	-0.092	0.108	0.771	0.369	-0.071	0.081	0.762	0.277
smooth ls	-0.173	0.21	0.829	0.821	-0.092	0.108	0.812	0.408	-0.071	0.081	0.806	0.302
smooth xavier	-0.173	0.21	0.794	0.75	-0.092	0.108	0.791	0.387	-0.071	0.081	0.788	0.289
	μ_X=1, σ_X^2=2 – $b_1 = b_2$=0.5											
standard kernel	-0.116	0.208	0.833	0.68	-0.057	0.104	0.835	0.342	-0.04	0.077	0.841	0.253

standard xavier	-0.116	0.208	0.84	0.703	-0.057	0.104	0.853	0.36	-0.04	0.077	0.862	0.267
smooth kernel	-0.137	0.204	0.835	0.717	-0.068	0.102	0.833	0.358	-0.048	0.075	0.839	0.266
smooth ls	-0.137	0.204	0.866	0.809	-0.068	0.102	0.864	0.395	-0.048	0.075	0.873	0.29
smooth xavier	-0.137	0.204	0.842	0.746	-0.068	0.102	0.85	0.379	-0.048	0.075	0.859	0.281
	μ_X=1, σ_X^2=1.666 – $b_1 = b_2$=0.4											
standard kernel	-0.091	0.205	0.855	0.673	-0.038	0.099	0.877	0.337	-0.026	0.073	0.887	0.248
standard xavier	-0.091	0.205	0.862	0.695	-0.038	0.099	0.894	0.354	-0.026	0.073	0.902	0.262
smooth kernel	-0.114	0.2	0.858	0.71	-0.049	0.098	0.877	0.351	-0.034	0.072	0.886	0.26
smooth ls	-0.114	0.2	0.885	0.792	-0.049	0.098	0.905	0.385	-0.034	0.072	0.908	0.28
smooth xavier	-0.114	0.2	0.865	0.737	-0.049	0.098	0.894	0.372	-0.034	0.072	0.901	0.273
	μ_X=1, σ_X^2=1.333 – $b_1 = b_2$=0.25											
standard kernel	-0.071	0.198	0.879	0.667	-0.025	0.095	0.904	0.332	-0.015	0.069	0.911	0.243
standard xavier	-0.071	0.198	0.886	0.691	-0.025	0.095	0.918	0.35	-0.015	0.069	0.923	0.255
smooth kernel	-0.095	0.193	0.884	0.704	-0.036	0.093	0.904	0.344	-0.023	0.068	0.913	0.256
smooth ls	-0.095	0.193	0.906	0.781	-0.036	0.093	0.923	0.373	-0.023	0.068	0.927	0.269
smooth xavier	-0.095	0.193	0.889	0.734	-0.036	0.093	0.918	0.364	-0.023	0.068	0.923	0.264
	μ_X=1, σ_X^2=1.111 – $b_1 = b_2$=0.1											
standard kernel	-0.054	0.196	0.889	0.666	-0.017	0.093	0.917	0.328	-0.01	0.066	0.924	0.239
standard xavier	-0.054	0.196	0.896	0.69	-0.017	0.093	0.928	0.344	-0.01	0.066	0.934	0.25
smooth kernel	-0.079	0.191	0.895	0.703	-0.027	0.091	0.917	0.341	-0.017	0.066	0.925	0.248
smooth ls	-0.079	0.191	0.913	0.772	-0.027	0.091	0.932	0.363	-0.017	0.066	0.939	0.26
smooth xavier	-0.079	0.191	0.899	0.733	-0.027	0.091	0.928	0.356	-0.017	0.066	0.936	0.257
	μ_X=1, σ_X^2=1.053 – $b_1 = b_2$=0.05											
standard kernel	-0.052	0.196	0.892	0.667	-0.015	0.092	0.916	0.327	-0.01	0.065	0.925	0.238
standard xavier	-0.052	0.196	0.898	0.691	-0.015	0.092	0.928	0.343	-0.01	0.065	0.935	0.248
smooth kernel	-0.076	0.191	0.897	0.703	-0.025	0.091	0.918	0.337	-0.016	0.065	0.928	0.246
smooth ls	-0.076	0.191	0.912	0.768	-0.025	0.091	0.934	0.36	-0.016	0.065	0.939	0.257
smooth xavier	-0.076	0.191	0.899	0.73	-0.025	0.091	0.928	0.354	-0.016	0.065	0.936	0.254
	μ_X=2, σ_X^2=20 – $b_1 = b_2$=0.95											
standard kernel	-1.178	0.308	0.012	0.802	-1.024	0.203	0.001	0.424	-0.969	0.174	0.001	0.319
standard xavier	-1.178	0.308	0.013	0.808	-1.024	0.203	0.002	0.432	-0.969	0.174	0.001	0.326
smooth kernel	-1.183	0.304	0.013	0.809	-1.028	0.2	0.002	0.437	-0.973	0.171	0.002	0.335
smooth ls	-1.183	0.304	0.039	1.001	-1.028	0.2	0.005	0.54	-0.973	0.171	0.002	0.416
smooth xavier	-1.183	0.304	0.014	0.82	-1.028	0.2	0.002	0.444	-0.973	0.171	0.001	0.339
	μ_X=2, σ_X^2=10 – $b_1 = b_2$=0.9											

standard kernel	-0.926	0.284	0.041	0.726	-0.743	0.187	0.009	0.384	-0.684	0.157	0.003	0.29
standard xavier	-0.926	0.284	0.044	0.736	-0.743	0.187	0.011	0.396	-0.684	0.157	0.004	0.302
smooth kernel	-0.937	0.279	0.042	0.744	-0.752	0.183	0.012	0.414	-0.691	0.153	0.006	0.312
smooth ls	-0.937	0.279	0.086	0.939	-0.752	0.183	0.023	0.51	-0.691	0.153	0.013	0.395
smooth xavier	-0.937	0.279	0.046	0.762	-0.752	0.183	0.012	0.42	-0.691	0.153	0.006	0.325
	μ_X=2, σ_X^2=4 – $b_1 = b_2$=0.75											
standard kernel	-0.581	0.269	0.163	0.618	-0.396	0.162	0.102	0.338	-0.336	0.134	0.083	0.262
standard xavier	-0.581	0.269	0.178	0.642	-0.396	0.162	0.12	0.364	-0.336	0.134	0.101	0.287
smooth kernel	-0.605	0.263	0.17	0.673	-0.412	0.158	0.11	0.38	-0.35	0.13	0.093	0.329
smooth ls	-0.605	0.263	0.286	0.9	-0.412	0.158	0.188	0.496	-0.35	0.13	0.163	0.39
smooth xavier	-0.605	0.263	0.19	0.709	-0.412	0.158	0.133	0.411	-0.35	0.13	0.116	0.328
	μ_X=2, σ_X^2=2.5 – $b_1 = b_2$=0.6											
standard kernel	-0.438	0.258	0.274	0.578	-0.261	0.158	0.261	0.334	-0.208	0.125	0.265	0.264
standard xavier	-0.438	0.258	0.299	0.613	-0.261	0.158	0.306	0.371	-0.208	0.125	0.31	0.297
smooth kernel	-0.469	0.252	0.286	0.657	-0.283	0.154	0.273	0.385	-0.225	0.122	0.278	0.313
smooth ls	-0.469	0.252	0.443	0.909	-0.283	0.154	0.403	0.51	-0.225	0.122	0.398	0.399
smooth xavier	-0.469	0.252	0.32	0.706	-0.283	0.154	0.322	0.43	-0.225	0.122	0.325	0.343
	μ_X=2, σ_X^2=2 – $b_1 = b_2$=0.5											
standard kernel	-0.376	0.261	0.337	0.569	-0.21	0.154	0.36	0.336	-0.165	0.12	0.372	0.268
standard xavier	-0.376	0.261	0.37	0.612	-0.21	0.154	0.411	0.379	-0.165	0.12	0.427	0.304
smooth kernel	-0.41	0.254	0.355	0.661	-0.232	0.15	0.374	0.398	-0.183	0.117	0.387	0.333
smooth ls	-0.41	0.254	0.522	0.926	-0.232	0.15	0.506	0.517	-0.183	0.117	0.51	0.401
smooth xavier	-0.41	0.254	0.397	0.717	-0.232	0.15	0.428	0.44	-0.183	0.117	0.441	0.351
	μ_X=2, σ_X^2=1.666 – $b_1 = b_2$=0.4											
standard kernel	-0.336	0.256	0.384	0.567	-0.174	0.152	0.443	0.341	-0.129	0.119	0.475	0.274
standard xavier	-0.336	0.256	0.422	0.616	-0.174	0.152	0.504	0.389	-0.129	0.119	0.534	0.313
smooth kernel	-0.373	0.25	0.406	0.668	-0.198	0.148	0.458	0.438	-0.148	0.116	0.485	0.325
smooth ls	-0.373	0.25	0.573	0.941	-0.198	0.148	0.596	0.524	-0.148	0.116	0.614	0.407
smooth xavier	-0.373	0.25	0.445	0.732	-0.198	0.148	0.518	0.451	-0.148	0.116	0.548	0.36
	μ_X=2, σ_X^2=1.333 – $b_1 = b_2$=0.25											
standard kernel	-0.288	0.262	0.438	0.568	-0.137	0.149	0.543	0.35	-0.096	0.115	0.589	0.281
standard xavier	-0.288	0.262	0.482	0.628	-0.137	0.149	0.602	0.402	-0.096	0.115	0.651	0.324
smooth kernel	-0.329	0.256	0.465	0.681	-0.162	0.145	0.559	0.408	-0.116	0.112	0.602	0.673
smooth ls	-0.329	0.256	0.628	0.966	-0.162	0.145	0.68	0.532	-0.116	0.112	0.715	0.411

smooth xavier	-0.329	0.256	0.508	0.755	-0.162	0.145	0.614	0.465	-0.116	0.112	0.662	0.37
	μ_X=2, σ_X^2=1.111 – $b_1 = b_2$=0.1											
standard kernel	-0.258	0.264	0.478	0.57	-0.108	0.15	0.613	0.359	-0.078	0.114	0.659	0.289
standard xavier	-0.258	0.264	0.522	0.639	-0.108	0.15	0.676	0.416	-0.078	0.114	0.723	0.334
smooth kernel	-0.301	0.257	0.512	0.691	-0.134	0.146	0.629	0.421	-0.097	0.11	0.672	0.354
smooth ls	-0.301	0.257	0.665	0.979	-0.134	0.146	0.742	0.539	-0.097	0.11	0.769	0.413
smooth xavier	-0.301	0.257	0.552	0.772	-0.134	0.146	0.685	0.476	-0.097	0.11	0.728	0.377
	μ_X=2, σ_X^2=1.053 – $b_1 = b_2$=0.05											
standard kernel	-0.248	0.257	0.496	0.574	-0.104	0.148	0.633	0.362	-0.073	0.113	0.682	0.291
standard xavier	-0.248	0.257	0.545	0.644	-0.104	0.148	0.695	0.42	-0.073	0.113	0.74	0.337
smooth kernel	-0.292	0.25	0.533	0.701	-0.13	0.144	0.648	0.431	-0.092	0.11	0.692	0.335
smooth ls	-0.292	0.25	0.686	0.988	-0.13	0.144	0.757	0.542	-0.092	0.11	0.783	0.412
smooth xavier	-0.292	0.25	0.574	0.782	-0.13	0.144	0.704	0.482	-0.092	0.11	0.744	0.379
	μ_X=3, σ_X^2=20 – $b_1 = b_2$=0.95											
standard kernel	-1.816	0.318	0.0	0.741	-1.589	0.222	0.0	0.394	-1.506	0.196	0.0	0.297
standard xavier	-1.816	0.318	0.0	0.746	-1.589	0.222	0.0	0.401	-1.506	0.196	0.0	0.304
smooth kernel	-1.822	0.313	0.0	0.75	-1.595	0.219	0.0	0.406	-1.511	0.193	0.001	0.311
smooth ls	-1.822	0.313	0.002	0.97	-1.595	0.219	0.0	0.533	-1.511	0.193	0.0	0.415
smooth xavier	-1.822	0.313	0.0	0.76	-1.595	0.219	0.0	0.416	-1.511	0.193	0.0	0.319
	μ_X=3, σ_X^2=10 – $b_1 = b_2$=0.9											
standard kernel	-1.484	0.309	0.002	0.63	-1.218	0.219	0.0	0.338	-1.124	0.191	0.0	0.258
standard xavier	-1.484	0.309	0.002	0.64	-1.218	0.219	0.0	0.352	-1.124	0.191	0.0	0.272
smooth kernel	-1.497	0.303	0.002	0.655	-1.228	0.215	0.001	0.365	-1.134	0.188	0.002	0.288
smooth ls	-1.497	0.303	0.013	0.924	-1.228	0.215	0.002	0.52	-1.134	0.188	0.002	0.412
smooth xavier	-1.497	0.303	0.003	0.675	-1.228	0.215	0.0	0.383	-1.134	0.188	0.0	0.302
	μ_X=3, σ_X^2=4 – $b_1 = b_2$=0.75											
standard kernel	-1.096	0.311	0.014	0.468	-0.795	0.221	0.009	0.281	-0.693	0.191	0.009	0.228
standard xavier	-1.096	0.311	0.017	0.492	-0.795	0.221	0.013	0.313	-0.693	0.191	0.015	0.259
smooth kernel	-1.121	0.304	0.018	0.542	-0.815	0.215	0.015	0.347	-0.711	0.186	0.018	0.3
smooth ls	-1.121	0.304	0.081	0.957	-0.815	0.215	0.051	0.578	-0.711	0.186	0.046	0.473
smooth xavier	-1.121	0.304	0.024	0.583	-0.815	0.215	0.019	0.382	-0.711	0.186	0.02	0.321
	μ_X=3, σ_X^2=2.5 – $b_1 = b_2$=0.6											
standard kernel	-0.929	0.331	0.031	0.409	-0.635	0.224	0.031	0.273	-0.537	0.196	0.033	0.234
standard xavier	-0.929	0.331	0.04	0.446	-0.635	0.224	0.044	0.317	-0.537	0.196	0.049	0.277
smooth kernel	-0.96	0.323	0.043	0.521	-0.659	0.218	0.046	0.365	-0.559	0.191	0.05	0.316

smooth ls	-0.96	0.323	0.168	1.041	-0.659	0.218	0.13	0.639	-0.559	0.191	0.127	0.529
smooth xavier	-0.96	0.323	0.058	0.575	-0.659	0.218	0.062	0.409	-0.559	0.191	0.066	0.357
	μ_X=3, σ^2_X=2 – $b_1 = b_2$=0.5											
standard kernel	-0.877	0.329	0.038	0.384	-0.569	0.23	0.051	0.275	-0.473	0.197	0.055	0.241
standard xavier	-0.877	0.329	0.047	0.426	-0.569	0.23	0.072	0.327	-0.473	0.197	0.08	0.29
smooth kernel	-0.91	0.321	0.051	0.512	-0.596	0.225	0.07	0.377	-0.497	0.191	0.081	0.337
smooth ls	-0.91	0.321	0.203	1.08	-0.596	0.225	0.189	0.677	-0.497	0.191	0.19	0.559
smooth xavier	-0.91	0.321	0.068	0.572	-0.596	0.225	0.094	0.431	-0.497	0.191	0.104	0.379
	μ_X=3, σ^2_X=1.666 – $b_1 = b_2$=0.4											
standard kernel	-0.832	0.335	0.046	0.37	-0.521	0.236	0.071	0.279	-0.425	0.2	0.082	0.248
standard xavier	-0.832	0.335	0.06	0.414	-0.521	0.236	0.1	0.337	-0.425	0.2	0.112	0.302
smooth kernel	-0.868	0.326	0.065	0.512	-0.55	0.229	0.1	0.388	-0.451	0.194	0.11	0.492
smooth ls	-0.868	0.326	0.241	1.119	-0.55	0.229	0.244	0.713	-0.451	0.194	0.253	0.583
smooth xavier	-0.868	0.326	0.089	0.578	-0.55	0.229	0.132	0.453	-0.451	0.194	0.142	0.397
	μ_X=3, σ^2_X=1.333 – $b_1 = b_2$=0.25											
standard kernel	-0.793	0.342	0.052	0.35	-0.472	0.239	0.096	0.285	-0.376	0.2	0.119	0.259
standard xavier	-0.793	0.342	0.067	0.399	-0.472	0.239	0.132	0.35	-0.376	0.2	0.159	0.32
smooth kernel	-0.831	0.334	0.074	0.511	-0.503	0.233	0.129	0.407	-0.404	0.195	0.155	1.885
smooth ls	-0.831	0.334	0.28	1.159	-0.503	0.233	0.307	0.745	-0.404	0.195	0.322	0.616
smooth xavier	-0.831	0.334	0.098	0.583	-0.503	0.233	0.166	0.474	-0.404	0.195	0.194	0.423
	μ_X=3, σ^2_X=1.111 – $b_1 = b_2$=0.1											
standard kernel	-0.754	0.348	0.062	0.341	-0.429	0.243	0.125	0.291	-0.34	0.204	0.155	0.271
standard xavier	-0.754	0.348	0.081	0.396	-0.429	0.243	0.167	0.361	-0.34	0.204	0.204	0.339
smooth kernel	-0.793	0.34	0.094	0.518	-0.462	0.236	0.167	0.422	-0.37	0.198	0.198	0.404
smooth ls	-0.793	0.34	0.324	1.221	-0.462	0.236	0.366	0.777	-0.37	0.198	0.389	0.646
smooth xavier	-0.793	0.34	0.124	0.598	-0.462	0.236	0.212	0.493	-0.37	0.198	0.248	0.449
	μ_X=3, σ^2_X=1.053 – $b_1 = b_2$=0.05											
standard kernel	-0.747	0.347	0.064	0.335	-0.429	0.244	0.127	0.294	-0.327	0.206	0.167	0.272
standard xavier	-0.747	0.347	0.084	0.391	-0.429	0.244	0.168	0.367	-0.327	0.206	0.222	0.34
smooth kernel	-0.788	0.339	0.095	0.515	-0.463	0.237	0.167	0.441	-0.357	0.2	0.216	0.408
smooth ls	-0.788	0.339	0.328	1.225	-0.463	0.237	0.369	0.783	-0.357	0.2	0.401	0.651
smooth xavier	-0.788	0.339	0.126	0.597	-0.463	0.237	0.212	0.5	-0.357	0.2	0.265	0.452

5.6 Proofs of the Main Results

Below, we use "$\lesssim$" to indicate an inequality up to universal constant. In most cases below, this means a constant independent of x and n. The floor function $\lfloor\cdot\rfloor$ takes as input a real number x, and gives as output the greatest integer less than or equal to x, denoted $\lfloor x\rfloor$. Similarly, the ceiling function $\lceil\cdot\rceil$ maps x to the least integer greater than or equal to x denoted $\lceil x\rceil$. We let $\mathrm{B}(\cdot,\cdot)$ denote the beta function, i.e., for all $a,b>0$, $\mathrm{B}(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}dt$. The identity function is denoted $\mathrm{I}: x\ni\mathbb{R}\mapsto x$.

5.6.1 Proof of Lemma 5.2.1

Observe that $\mathbb{E}[|Y|]<\infty$ implies $tS_Y(t)\to 0$ and $tF_Y(-t)\to 0$ as $t\to\infty$. Thus $\mathbb{E}[|Y|^p]<\infty$ implies $t^pS_Y(t)\to 0$ and $t^pF_Y(-t)\to 0$ as $t\to\infty$. The convergence to 0 of $t^pS_Y(t)$ implies that there exists $C>0$ and t_1 such that for all $t\geq t_1$,

$$|t|^p(1-F_Y(t))\leq C.$$

This implies that for all $u\geq F_Y(t_1)$, $|F_Y^{-1}(u)|^p(1-u)\leq C$ or, equivalently,

$$|F_Y^{-1}(u)|\leq C(1-u)^{-1/p}.$$

Hence, there exists $C_1>0$ such that for $u\geq F_Y(t_1)$,

$$|F_Y^{-1}(u)|\leq C_1[u(1-u)]^{-1/p}.$$

Using $t^pF_Y(-t)\to 0$ and a similar reasoning, there exists C_2 and $t_2\leq t_1$ such that for all $u\leq t_2$, $|F_Y^{-1}(u)|\leq C_2[u(1-u)]^{-1/p}$. The result follows since $|F_Y^{-1}(u)[u(1-u)]^{1/p}|$ is bounded on $[t_2,t_1]$. □

5.6.2 Proof of Theorem 5.2.2

Consider the following decomposition:

$$\widehat{\theta}_{n_1,n_2}-\theta_0=\underbrace{\int_0^1 F_Y^{-1}\circ\mathbb{G}_{n_1}^{-1}\,dF_U-\int_0^1 F_Y^{-1}\,dF_U}_{:=T_{1n_1}}+\underbrace{\int_0^1 F_Y^{-1}\circ\mathbb{G}_{n_1}^{-1}\,d\widehat{F}_U-\int_0^1 F_Y^{-1}\circ\mathbb{G}_{n_1}^{-1}\,dF_U}_{:=T_{2N}}.$$

The proof proceeds in three steps. In the first step, we prove that $\sqrt{N}T_{1n_1}$ is linear up to a negligible remainder term. In the second step, we prove the same result for $\sqrt{N}T_{2N}$. The last step concludes.

First step: linearization of $\sqrt{N}T_{1n_1}$. By Lemma 5.7.1 followed by an integration by part,

$$\begin{aligned}\sqrt{N}T_{1n_1} &= \sqrt{n}\left[\int_0^1 F_Y^{-1}\,dF_U\circ\mathbb{G}_{n_1} - \int_0^1 F_Y^{-1}\,dF_U\right]\\ &= \sqrt{N}\left\{\int_{\xi_{(1)}}^{\xi_{(n_1)}} F_Y^{-1}\,d\left[F_U\circ\mathbb{G}_{n_1} - F_U\right] - \int_0^{\xi_{(1)}} F_Y^{-1}dF_U - \int_{\xi_{(n_1)}}^1 F_Y^{-1}dF_U\right\}\\ &= -\sqrt{N}\left\{\int_{\xi_{(1)}}^{\xi_{(n_1)}} \left[F_U\circ\mathbb{G}_{n_1} - F_U\right]dF_Y^{-1} - \int_0^{\xi_{(1)}} F_Y^{-1}dF_U - \int_{\xi_{(n_1)}}^1 F_Y^{-1}dF_U\right\},\end{aligned}$$

where the last equality relies on Assumption 5.2.2, $d_1+b_1<1$ and $d_2+b_2<1$. Next, using Assumption 5.2.2 again,

$$\begin{aligned}\left|\int_0^{\xi_{(1)}} F_Y^{-1}dF_U\right| &\lesssim \mathbb{1}\{\xi_{(1)}\geq 1/2\}\left|\int_0^1 F_Y^{-1}dF_U\right| + \mathbb{1}\{\xi_{(1)}<1/2\}\int_0^{\xi_{(1)}} t^{-b_1-d_1}dt\\ &\lesssim \mathbb{1}\{\xi_{(1)}\geq 1/2\} + \xi_{(1)}^{1-b_1-d_1}.\end{aligned}$$

Thus, because $\xi_{(1)}=O_p(1/n_1)$ and $b_1+d_1<1/2$, $\sqrt{N}\int_0^{\xi_{(1)}} F_Y^{-1}dF_U = o_p(1)$. Similarly, $\sqrt{N}\int_{\xi_{(n_1)}}^1 F_Y^{-1}dF_U = o_p(1)$. Hence,

$$\sqrt{N}T_{1n_1} = -\sqrt{N}\int_{\xi_{(1)}}^{\xi_{(n_1)}}[\mathbb{G}_{n_1} - I]d\Lambda + R_N + o_p(1),$$

where Λ is the measure defined by $d\Lambda/dF_Y^{-1} = f_U$ and

$$R_N := \sqrt{N}\left(\int_{\xi_{(1)}}^{\xi_{(n_1)}}[\mathbb{G}_{n_1} - \mathrm{I}]f_U\,dF_Y^{-1} - \int_{\xi_{(1)}}^{\xi_{(n_1)}}\left[F_U\circ\mathbb{G}_{n_1} - F_U\right]dF_Y^{-1}\right).$$

We show below that $R_N=o_p(1)$, which further proves that

$$\sqrt{N}T_{1n_1} = -\sqrt{N}\int_{\xi_{(1)}}^{\xi_{(n_1)}}[\mathbb{G}_{n_1} - \mathrm{I}]d\Lambda + o_p(1).$$

By the mean value theorem, there exists $T_{n_1}(t)\in(\mathbb{G}_{n_1}(t),t)$ such that

$$R_N = \sqrt{N}\int_{\xi_{(1)}}^{\xi_{(n_1)}} \underbrace{[f_U - f_U\circ T_{n_1}]}_{:=A_{n_1}}[\mathbb{G}_{n_1} - \mathrm{I}]\;dF_Y^{-1}.$$

By Assumption 5.2.2, there exists $\delta>0$ such that $b_j+d_j<1/2-\delta$. Further, let $\delta_j>0$ be such that

$$b_j + d_j < 1/2 - \delta - \delta_j. \tag{5.6.1}$$

Then let $q(t) = t^{1/2-\delta_1}(1-t)^{1/2-\delta_2}$. From what precedes, we have

$$|R_N| \quad \leq \sqrt{\frac{N}{n_1}}\sup_{t\in(0,1)}\left|\frac{\sqrt{n_1}(\mathbb{G}_{n_1}(t)-t)}{q(t)}\right|\int_{\xi_{(1)}}^{\xi_{(n_1)}}|A_{n_1}(t)|q(t)\,dF_Y^{-1}(t). \tag{5.6.2}$$

We now show that the last term tends to 0 almost surely. First, by convergence of $\mathbb{G}_{n_1}(t)$ to t, we have, for all $t \in (0,1)$, $T_{n_1}(t) \xrightarrow{\text{a.s.}} t$. Then, by continuity of f_U, $A_{n_1}(t) \xrightarrow{\text{a.s.}} 0$ for all $t \in (0,1)$. Fix $\varepsilon > 0$. By Theorem 10.6.1 in Shorack and Wellner (1986), we have, for all $t \geq \xi_{(1)}$ and all n_1 large enough,

$$\mathbb{G}_{n_1}(t) \leq (1+\varepsilon)t^{1-\delta/2} \leq (1+\varepsilon)t^{1-\delta}.$$

Now, let $B(t) := C_U t^{-b_1}(1-t)^{-b_2}$. Then, by Assumption 5.2.2 and because B is a convex function, we obtain, for all $t \in [\xi_{(1)}, \xi_{(n_1)}]$,

$$\begin{aligned} |A_{n_1}(t)| &\leq [B(\mathbb{G}_{n_1}(t)) \vee B(t)] + B(t) \\ &\lesssim t^{-b_1-\delta}(1-t)^{-b_2-\delta}, \quad \text{a.s.} \end{aligned}$$

Therefore,

$$|A_{n_1}(t)|\mathbb{1}\{t \in [\xi_{(1)}, \xi_{(n_1)}]\}q(t) \lesssim t^{1/2-b_1-\delta-\delta_1}(1-t)^{1/2-b_2-\delta-\delta_2}.$$

Moreover, by (5.6.1) and Lemma 5.7.2,

$$\int_0^1 t^{1/2-b_1-\delta-\delta_1}(1-t)^{1/2-b_2-\delta-\delta_2}\, dF_Y^{-1}(t) < \infty.$$

Then, by the dominated convergence theorem,

$$\int_{\xi_{(1)}}^{\xi_{(n_1)}} |A_{n_1}(t)|q(t)\, dF_Y^{-1}(t) \xrightarrow{\text{a.s.}} 0. \tag{5.6.3}$$

Next, by Equation (2) in Chapter 2, Section 7 (page 141) in Shorack and Wellner (1986), there exists a Brownian bridge $\mathbb{U}$ such that

$$\sup_{t\in(0,1)} \left| \frac{\sqrt{n_1}(\mathbb{G}_{n_1}(t) - t) - \mathbb{U}(t)}{q(t)} \right| = o_P(1).$$

Hence, since $||./q||$ is a norm, the triangular inequality yields

$$\sup_{t\in(0,1)} \left| \frac{\sqrt{n_1}(\mathbb{G}_{n_1}(t) - t)}{q(t)} \right| \leq \sup_{t\in(0,1)} |\mathbb{U}(t)/q(t)| + o_p(1) = O_p(1) + o_p(1) = O_p(1),$$

by inequality (17) p.451 in Shorack and Wellner (1986) with $a = 0, b = 1/2$, and by noticing that $\mathbb{U}/q$ has the same distribution on $[0, 1/2]$ and $[1/2, 1]$, and that the integral on the right-hand side of the inequality is finite for $q(t) = [t(1-t)]^a$ with $a < 1/2$. This, together with (5.6.2) and (5.6.3), implies that $R_N = o_p(1)$. Let

$$R'_N = \sqrt{N}\left(\int_{\xi_{(1)}}^{\xi_{(n_1)}} [\mathbb{G}_{n_1} - I]d\Lambda - \int_0^1 [\mathbb{G}_{n_1} - I]d\Lambda \right).$$

The last step is to show that $R'_N \xrightarrow{p} 0$, which further proves, provided that $\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\eta_i = O_p(1)$ (which we show below),

$$\sqrt{N}T_{1n_1} = \frac{\lambda_1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\eta_i + o_p(1), \tag{5.6.4}$$

with $\eta_i := -\int_0^1[\mathbb{1}\{F_Y(Y_i)\le t\} - t]d\Lambda(t)$. We have

$$|R'_N| \le \sqrt{N}\int_0^{\xi_{(1)}} x d\Lambda(x) + \sqrt{N}\int_{\xi_{(n_1)}}^1 (1-x)d\Lambda(x).$$

Consider the first term, the second can be handled similarly. Let $\delta \in (b_1+d_1, 1/2)$. With probability tending to 1, $\xi_{(1)} \le 1/2$. Moreover, under this event, we have, using Assumption 5.2.2 again,

$$\begin{aligned}\sqrt{N}\int_0^{\xi_{(1)}} x d\Lambda(x) &\lesssim \sqrt{N}\xi_{(1)}^{1-\delta}\int_0^{1/2} x^\delta d\Lambda(x)\\ &\lesssim \sqrt{N}\xi_{(1)}^{1-\delta}\int_0^{1/2} x^{\delta-b_1} dF_Y^{-1}(x).\end{aligned}$$

By Lemma 5.7.2, $\int_0^{1/2} x^{\delta-b_1}dF_Y^{-1}(x) < \infty$. Moreover, because $\xi_{(1)} = O_p(1/n_1)$, we also have $\sqrt{N}\xi_{(1)}^{1-\delta} = o_p(1)$. The result follows.

Second step: linearization of $\sqrt{N}T_{2N}$. By Lemma 5.7.1 followed by an integration by part,

$$\begin{aligned}\sqrt{N}T_{2N} &= \sqrt{N}\int_0^1 F_Y^{-1}\, d\left[\widehat{F}_U\circ\mathbb{G}_{n_1} - F_U\circ\mathbb{G}_{n_1}\right]\\ &= \sqrt{N}\left[F_Y^{-1}(t)\left(\widehat{F}_U(\mathbb{G}_{n_1}(t)) - F_U(\mathbb{G}_{n_1}(t))\right)\right]_0^1\\ &\quad - \sqrt{N}\int_0^1\left[\widehat{F}_U\circ\mathbb{G}_{n_1} - F_U\circ\mathbb{G}_{n_1}\right] dF_Y^{-1}\\ &= -\sqrt{N}\int_0^1\left[\widehat{F}_U\circ\mathbb{G}_{n_1} - F_U\circ\mathbb{G}_{n_1}\right] dF_Y^{-1},\end{aligned} \tag{5.6.5}$$

since for $t\in(0,\xi_{(1)})$, $\mathbb{G}_{n_1}(t) = 0$ and $\widehat{F}_U(0) = F_U(0) = 0$ because $(U_i)_{i=1,\dots,n_2}$ is an iid sample of random variables absolutely continuous with respect to the Lebesgue measure on $[0,1]$. Symmetrically, for $t\in(\xi_{(n_1)},1)$, $\mathbb{G}_{n_1}(t) = 1$ and $\widehat{F}_U(1) = F_U(1) = 1$. We now prove that

$$-\sqrt{N}\int_0^1\left[\widehat{F}_U\circ\mathbb{G}_{n_1} - F_U\circ\mathbb{G}_{n_1}\right] dF_Y^{-1} = -\sqrt{N}\int_0^1\left[\widehat{F}_U - F_U\right] dF_Y^{-1} + o_p(1). \tag{5.6.6}$$

Let $\mathbb{V}_{n_2} = \sqrt{n_2}(\widehat{F}_U\circ F_U^{-1} - \mathrm{I})$ denote the empirical process associated with the uniform variables $(F_U(U_i))_{i=1,\dots,n}$ and define

$$R_N = \int_0^1 (\mathbb{V}_{n_2}\circ F_U\circ\mathbb{G}_{n_1} - \mathbb{V}_{n_2}\circ F_U)\, dF_Y^{-1}.$$

Equation (5.6.6) is equivalent to $\sqrt{N/n_2}R_N = o_p(1)$. We actually prove the stronger result that $\mathbb{E}[|R_N|] \to 0$. For that purpose, let $I_{n_1}(x) = (x, \mathbb{G}_{n_1}(x)]$ if $\mathbb{G}_{n_1}(x) > x$, $I_{n_1}(x) = [\mathbb{G}_{n_1}(x), x)$ if $\mathbb{G}_{n_1}(x) < x$ and $\emptyset$ otherwise. Finally, let $S_{n_1}(x) = \operatorname{sgn}(\mathbb{G}_{n_1}(x) - x)$. Observe first that

$$\mathbb{V}_{n_2} \circ F_U \circ \mathbb{G}_{n_1}(x) - \mathbb{V}_{n_2} \circ F_U(x) = S_n(x) Z_n(x), \tag{5.6.7}$$

with

$$Z_N(x) = \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n_2} \left[\mathbb{1}\{U_i \in I_{n_1}(x)\} - P_U(I_{n_1}(x)) \right],$$

where $P_U([a,b]) = P_U((a,b]) = P_U([a,b)) = P_U((a,b)) = F_U(a) - F_U(b)$ for all $(a,b) \in [0,1], a \leq b$. Then,

$$\begin{aligned}
\mathbb{E}\left[|R_N| \, |(\xi_i)_i\right] &\leq \mathbb{E}\left[\int_0^1 |\mathbb{V}_{n_2} \circ F_U \circ \mathbb{G}_{n_1} - \mathbb{V}_{n_2} \circ F_U| \; dF_Y^{-1} \middle| (\xi_i)_i \right] \\
&= \int_0^1 \mathbb{E}\left[|\mathbb{V}_{n_2} \circ F_U \circ \mathbb{G}_{n_1} - \mathbb{V}_{n_2} \circ F_U| \; |(\xi_i)_i \right] dF_Y^{-1} \\
&\leq \int_0^1 \mathbb{E}\left[Z_N(x)^2 | (\xi_i)_i \right]^{1/2} dF_Y^{-1}(x) \\
&= \int_0^1 \mathbb{V}\left[\mathbb{1}\{U_1 \in I_{n_1}(x)\} | (\xi_i)_i \right]^{1/2} dF_Y^{-1}(x) \\
&\leq \int_0^1 |P_U(I_{n_1}(x))|^{1/2} \, dF_Y^{-1}(x).
\end{aligned} \tag{5.6.8}$$

The first equality follows by Fubini-Tonelli's theorem, the second inequality uses (5.6.7) and the Cauchy-Schwarz inequality and the second equality holds since conditional on the $(\xi_i)_i$, the variables $\mathbb{1}\{U_i \in I_{n_1}(x)\} - P_U(I_{n_1}(x))$ are i.i.d. with mean zero. As a result,

$$\begin{aligned}
\mathbb{E}\left[|R_N|\right] &\leq \int_0^1 \mathbb{E}\left[|P_U(I_{n_1}(x)|^{1/2} \right] dF_Y^{-1}(x) \\
&\leq \int_0^1 \mathbb{E}\left[|P_U(I_{n_1}(x)|] \right]^{1/2} dF_Y^{-1}(x),
\end{aligned} \tag{5.6.9}$$

where the first inequality follows by (5.6.8) and Fubini-Tonelli's theorem, whereas the second is due to Jensen's inequality. Now, by the law of large numbers and the continuous mapping theorem, $|F_U(\mathbb{G}_{n_1}(x)) - F_U(x)| \xrightarrow{p} 0$ for all $x \in [0,1]$. Moreover, $|F_U(\mathbb{G}_{n_1}(x)) - F_U(x)| \leq 1$. Hence, for all $x \in [0,1]$,

$$\mathbb{E}\left[|F_U(x) - F_U(\mathbb{G}_{n_1}(x))|\right] \to 0.$$

We now apply the dominated convergence theorem to prove that $\mathbb{E}[|R_N|] \to 0$. Because $x \mapsto \mathbb{E}[|P_U(I_{n_1}(x))|]^{1/2}$ is bounded by 1 for all n_1, it is actually enough to bound this function for x close to 0 and close to 1. Also, by symmetry, we can focus without loss of generality on the neighborhood of 0. We prove that

$$\mathbb{E}\left[|P_U(I_{n_1}(x))|\right] \lesssim x^{1-b_1}. \tag{5.6.10}$$

Then the result follows by Lemma 5.7.2 combined with Assumption 5.2.2. To prove (5.6.10), we apply Lemma 5.7.4 with $Q_n(x) := \mathbb{G}_{n_1}(x)$ and $\delta < \exp(-1)/2$. If $x \geq 1/n_1$, Cauchy-Schwarz inequality yields

$$\mathbb{E}\left[|\mathbb{G}_{n_1}(x) - x|\right] \leq \left[\frac{x(1-x)}{n_1}\right]^{1/2} \leq 2x, \tag{5.6.11}$$

since $n_1^{1/2} \geq x^{-1/2}$. If $x < 1/n_1$, (5.6.11) holds as well by Theorem 1 in Berend and Kontorovich (2013). Hence, (5.6.11) holds for all $x \in (0, \bar{\delta}/2)$. Next, let $n_0 \in \mathbb{N}$, $n_0 \geq 4/(1-\bar{\delta})^2$. By Kiefer's inequality (see, e.g. van der Vaart and Wellner, 1996, Corollary A.6.3), we have, for all $x \in [0, \delta]$ and all $n_1 \geq n_0$,

$$\mathbb{P}(\mathbb{G}_{n_1}(x) > 1/2) \leq (ex)^{n_1(1-\bar{\delta})^2/4} \lesssim x. \tag{5.6.12}$$

Thus, we can apply Lemma 5.7.4, which yields (5.6.10).

Hence, (5.6.6) holds. Combined with (5.6.5), this implies, provided that $\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2} \varepsilon_i = O_p(1)$ (which we show below),

$$\sqrt{N} T_{2N} = \frac{\lambda_2}{\sqrt{n_2}} \sum_{i=1}^{n_2} \varepsilon_i + o_p(1), \tag{5.6.13}$$

with $\varepsilon_i = -\int_0^1 [\mathbb{1}\{U_i \leq t\} - F_U(t)]\, dF_Y^{-1}(t)$.

Third step: conclusion. By definition of η_i and ε_i, we have $\mathbb{E}[\eta_i] = \mathbb{E}[\varepsilon_i] = 0$ and

$$\mathbb{E}[\eta_i^2] = \int_0^1 \int_0^1 (s \wedge t - st) f_U(s) f_U(t)\, dF_Y^{-1}(s)\, dF_Y^{-1}(t),$$
$$\mathbb{E}[\varepsilon_i^2] = \int_0^1 \int_0^1 (F_U(s \wedge t) - F_U(s) F_U(t))\, dF_Y^{-1}(s)\, dF_Y^{-1}(t).$$

Moreover, under Assumption 5.2.1, η_i and ε_i are independent. The result follows by the central limit theorem. □

5.6.3 Proof of Theorem 5.3.1

We first decompose the difference $\check{\theta}_{n_1,n_2,n_3} - \theta_0$ into three parts that we study independently:

$$\check{\theta}_{n_1,n_2,n_3} - \theta_0 = \underbrace{\int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}\, dF_U - \int_0^1 F_Y^{-1}\, dF_U}_{=T_{1n_1}} + \underbrace{\int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}\, d\widehat{F}_U - \int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}\, dF_U}_{=T_{2N}}$$
$$+ \underbrace{\int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U - \int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}\, d\widehat{F}_U}_{:=T_{3N}}.$$

This decomposition is convenient as T_{1n_1} and T_{2N} have already been analyzed in the proof of Theorem 5.2.2. We then prove the result in eight steps. We first show that

$$\sqrt{N}T_{3N} = -\sqrt{N}\int_0^1 \left[\widehat{F}_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1} - \widehat{F}_U \circ \mathbb{G}_{n_1}\right] dF_Y^{-1}. \tag{5.6.14}$$

where $\overline{\mathbb{G}}_{n_1}$ is defined below. Second, we show that

$$\sqrt{N}T_{3N} = -\sqrt{N}\underbrace{\int_0^1 \left[F_U \circ \mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1} - F_U \circ \mathbb{G}_{n_1}\right] dF_Y^{-1}}_{:=J_{1N}} + o_p(1). \tag{5.6.15}$$

Let us then write $-\sqrt{N}J_{1N} = \sqrt{N}J_{2n_3} + R_{1N} + R_{2N} + R_{3N} + R_{4N}$, with:

$$J_{2n_3} := -\int_{1/n_3}^{1-1/n_3} \left[\mathbb{H}_{n_3}^{-1}(x) - \mathbb{E}[\mathbb{H}_{n_3}^{-1}(x)]\right] f_U(x)\, dF_Y^{-1}(x), \tag{5.6.16}$$

$$R_{1N} := -\sqrt{N}\left(J_{1N} - \int_{\xi_{(1)}}^{\xi_{(n_1)}} \left[\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1} - \mathbb{G}_{n_1}\right] f_U\, dF_Y^{-1}\right), \tag{5.6.17}$$

$$R_{2N} := -\sqrt{N}\left(\int_{\xi_{(1)}}^{\xi_{(n_1)}} \left[\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1} - \mathbb{G}_{n_1}\right] f_U\, dF_Y^{-1} - \int_{\xi_{(1)}}^{\xi_{(n_1)}} \left[\mathbb{H}_{n_3}^{-1} - \mathrm{I}\right] f_U\, dF_Y^{-1}\right), \tag{5.6.18}$$

$$R_{3N} := \sqrt{N}\int_{\xi_{(1)}}^{\xi_{(n_1)}} \left[x - \mathbb{E}[\mathbb{H}_{n_3}^{-1}(x)]\right] f_U(x)\, dF_Y^{-1}(x), \tag{5.6.19}$$

$$\begin{aligned} R_{4N} := \sqrt{N}\Bigg(&\int_{1/n_3}^{1-1/n_3} \left[\mathbb{H}_{n_3}^{-1}(x) - \mathbb{E}[\mathbb{H}_{n_3}^{-1}(x)]\right] f_U(x)\, dF_Y^{-1}(x) \\ &- \int_{\xi_{(1)}}^{\xi_{(n_1)}} \left[\mathbb{H}_{n_3}^{-1}(x) - \mathbb{E}[\mathbb{H}_{n_3}^{-1}(x)]\right] f_U(x)\, dF_Y^{-1}(x)\Bigg). \end{aligned} \tag{5.6.20}$$

In the third to sixth steps, we prove that each of the four terms $R_{1N} - R_{4N}$ tends to 0 in probability. In the seventh step, we show that $\sqrt{N}J_{2n_3}$ tends to a normal distribution. The eighth step concludes.

First step: Equation (5.6.14) **holds.** Let $X_{n_3}^0 := [0, \zeta_{(1)}]$ and $X_{n_3}^1 := [\zeta_{(n_3)}, 1]$. For all $t \in [0,1]$, let us also define

$$\begin{aligned} \overline{\mathbb{G}}_{n_1}(t) &= \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{1}\{\xi_i \le t\} + \frac{1}{n_1}\sum_{i=1}^{n_1-1} \mathbb{1}\{\xi_i < t < \xi_{i+1}\} \\ &= \mathbb{G}_{n_1}(t) + \frac{1}{n_1}\sum_{i=1}^{n_1-1} \mathbb{1}\{\xi_i < t < \xi_{i+1}\}. \end{aligned}$$

Then, remark that $\mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}$ is the generalized inverse of $\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}$. Then, by splitting the first integral in $\sqrt{N}T_{3N}$ and applying Lemma 5.7.1, we obtain

$$\begin{aligned}
\sqrt{N}T_{3N} &= \sqrt{N}\left(\int_{(X_{n_3}^0 \cup X_{n_3}^1)^c} F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U - \int_0^1 F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1}\, d\widehat{F}_U \right.\\
&\quad \left. + \int_{X_{n_3}^0 \cup X_{n_3}^1} F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U \right)\\
&= \sqrt{N}\left(\int_{\xi_{(1)}}^{\xi_{(n_1)}} F_Y^{-1}\, d\left[\widehat{F}_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}\right] - \int_0^1 F_Y^{-1}\, d\left[\widehat{F}_U \circ \mathbb{G}_{n_1}\right] \right.\\
&\quad \left. + \int_{X_{n_3}^0 \cup X_{n_3}^1} F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U \right)\\
&= \sqrt{N}\int_{\xi_{(1)}}^{\xi_{(n_1)}} F_Y^{-1}\, d\left[\widehat{F}_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1} - \widehat{F}_U \circ \mathbb{G}_{n_1}\right]\\
&\quad + \sqrt{N}\int_{X_{n_3}^0 \cup X_{n_3}^1} F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U, \qquad (5.6.21)
\end{aligned}$$

where we used the fact that $\widehat{F}_U \circ \mathbb{G}_{n_1}$ is constant on the two segments $[0, \xi_{(1)}]$ and $[\xi_{(n_1)}, 1]$ to obtain the third equality. Remark that

$$\begin{aligned}
&\sqrt{N}\left[F_Y^{-1}(t)\left(\widehat{F}_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_n(t) - \widehat{F}_U \circ \mathbb{G}_{n_1}(t)\right)\right]_{t=\xi_{(1)}}^{t=\xi_{(n_1)}}\\
&\quad = \sqrt{N}\Big[\mathbb{1}\left\{\zeta_{(n_3)} < U_{(n_2)}\right\} F_Y^{-1}(\xi_{(n_1)})\left(\widehat{F}_U(\zeta_{(n_3)}) - 1\right) - \mathbb{1}\left\{\zeta_{(1)} \geq U_{(1)}\right\} F_Y^{-1}(\xi_{(1)})\widehat{F}_U(\zeta_{(1)})\Big].
\end{aligned}$$

Also, since $\mathbb{H}_{n_3}$ is constant on the two segments $X_{n_3}^0$ and $X_{n_3}^1$, we have

$$\begin{aligned}
&\sqrt{N}\int_{X_n^0 \cup X_{n_3}^1} F_Y^{-1} \circ \mathbb{G}_{n_1}^{-1} \circ \mathbb{H}_{n_3}\, d\widehat{F}_U\\
&\quad = \sqrt{N}\left[\widehat{F}_U(1) - \widehat{F}_U(\zeta_{(n_3)})\right] F_Y^{-1}(\xi_{(n_1)}) + \sqrt{n}\left[\widehat{F}_U(\zeta_{(1)}) - \widehat{F}_U(0)\right] F_Y^{-1}(\xi_{(1)})\\
&\quad = \sqrt{N}\Big[\mathbb{1}\left\{\zeta_{(1)} \geq U_{(1)}\right\} F_Y^{-1}(\xi_{(1)})\widehat{F}_U(\zeta_{(1)}) - \mathbb{1}\left\{\zeta_{(n_3)} < U_{(n_2)}\right\} F_Y^{-1}(\xi_{(n_1)})\left(\widehat{F}_U(\zeta_{(n_3)}) - 1\right)\Big].
\end{aligned}$$

Thus, an integration by part of the first term in (5.6.21) yields (5.6.14).

Second step: Equation (5.6.15) **holds.** From (5.6.14), we have

$$\begin{aligned}
\sqrt{N}T_{3N} \;=& -\sqrt{N}\underbrace{\int_0^1 \left[F_U \circ \mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1} - F_U \circ \mathbb{G}_{n_1}\right] dF_Y^{-1}}_{=:J_{1N}}\\
&- \sqrt{N}\int_0^1 \left[\widehat{F}_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1} - F_U \circ \mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}\right] dF_Y^{-1}\\
&- \sqrt{N}\int_0^1 \left[F_U \circ \mathbb{G}_{n_1} - \widehat{F}_U \circ \mathbb{G}_{n_1}\right] dF_Y^{-1}.
\end{aligned}$$

We show below that

$$\sqrt{N}\int_0^1 \left[\widehat{F}_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1} - F_U \circ \mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}\right] dF_Y^{-1} = \sqrt{N}\int_0^1 \left[\widehat{F}_U - F_U\right] dF_Y^{-1} + o_p(1). \tag{5.6.22}$$

Once combined with (5.6.6), this proves (5.6.15). To prove (5.6.22), we follow closely the proof of (5.6.6). Recall that $\mathbb{V}_{n_2} = \sqrt{n_2}(\widehat{F}_U \circ F_U^{-1} - \mathrm{I})$, and let

$$R_N = \int_0^1 \left(\mathbb{V}_{n_2} \circ F_U \circ \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1} - \mathbb{V}_{n_2} \circ F_U\right) dF_Y^{-1},$$

and $\bar{I}_N(x) = (x, \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x)]$ if $\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) > x$, $\bar{I}_N(x) = [\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x), x)$ if $\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) < x$ and $\emptyset$ otherwise. We prove that $\mathbb{E}[|R_N|] \to 0$. Reasoning as to obtain (5.6.9) (but conditioning first on $(\xi_i, \zeta_i)_i$ instead of just on $(\xi_i)_i$), we get

$$\mathbb{E}\left[|R_N|\right] \le \int_0^1 \mathbb{E}\left[\left|F_U(x) - F_U(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x))\right|\right]^{1/2} dF_Y^{-1}(x).$$

Because $\overline{\mathbb{G}}_{n_1}(x) \xrightarrow{p} x$, by uniform convergence of $\mathbb{H}_{n_3}^{-1}$ towards I and the continuous mapping theorem, $|F_U(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x)) - F_U(x)| \xrightarrow{p} 0$ for all $x \in [0,1]$. Moreover, $|F_U(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x)) - F_U(x)| \le 1$. Hence, for all $x \in [0,1]$,

$$\mathbb{E}\left[\left|F_U(x) - F_U(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x))\right|\right] \to 0.$$

Next, we show $\mathbb{E}[|R_N|] \to 0$ by proving

$$\mathbb{E}\left[\left|F_U(x) - F_U(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x))\right|\right] \lesssim x^{1-b_1}. \tag{5.6.23}$$

and applying the dominated convergence theorem. As in Theorem 5.2.2, we apply Lemma 5.7.4 with $Q_n(x) := \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x)$. The two conditions of this lemma are checked in Lemma 5.7.3. Hence, (5.6.23), and thus (5.6.15), hold.

Third step: $R_{1N} = o_p(1)$. Recall that R_{1N} is defined in (5.6.17). Let $j_N := \lceil n_1/n_3 \rceil \ge 1$. We split R_{1N} into two integrals: $R_{1N,1}$ is obtained by integrating on the segment $[\xi_{(1)}, \xi_{(j_N)}]$, and $R_{1N,2}$ is obtained by integrating on the segment $[\xi_{(j_N)}, \xi_{(n_1)}]$. Let us first focus on $R_{1N,2}$. By the mean value theorem, for all $t \in [\xi_{(j_N)}, \xi_{(n_1)}]$, there exists $T_N(t) \in (\mathbb{G}_{n_1}(t), \mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}(t))$ such that

$$R_{1N,2} = \sqrt{\frac{N}{n_3}}\sqrt{n_3}\int_{\xi_{(j_N)}}^{\xi_{(n_1)}} \underbrace{[f_U - f_U \circ T_N]}_{:=A_N}\left[\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1} - \mathbb{G}_{n_1}\right] dF_Y^{-1}.$$

By Assumption 5.2.2, there exists $\delta > 0$ such that $b_j + d_j < 1/2 - \delta$. Further, let $\delta_j > 0$ be such that

$$b_j + d_j < 1/2 - \delta - \delta_j. \tag{5.6.24}$$

Then let $q(t) = t^{1/2-\delta_1}(1-t)^{1/2-\delta_2}$. From what precedes, we have

$$|R_{1N,2}| \quad \leq \sqrt{\frac{N}{n_3}} \sup_{t\in(1/n_3,1-1/n_1)} \left| \frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)} \right| \int_{\xi_{(j_N)}}^{\xi_{(n_1)}} |A_N(t)| q(t)\, dF_Y^{-1}(t). \tag{5.6.25}$$

We now show that the integral in (5.6.25) tends to 0 almost surely. First, by uniform convergence of $\mathbb{H}_{n_3}^{-1}$ towards I and convergence of $\mathbb{G}_{n_1}(t)$ to t, we have, for all $t \in (0,1)$, $T_N(t) \xrightarrow{\text{a.s.}} t$. Then, by continuity of f_U, $A_N(t) \xrightarrow{\text{a.s.}} 0$ for all $t \in (0,1)$. Fix $\varepsilon > 0$. By Theorem 10.6.1 in Shorack and Wellner (1986), we have, for all $t \geq \xi_{(1)}$ and all n_1 large enough,

$$\mathbb{G}_{n_1}(t) \leq (1+\varepsilon)t^{1-\delta/2} \leq (1+\varepsilon)t^{1-\delta}.$$

Moreover, by the same theorem, we have, for all $u \geq 1/n_3$,

$$\mathbb{H}_{n_3}^{-1}(u) \leq (1+\varepsilon)u^{(1-\delta/2)}.$$

Then, since $\mathbb{G}_{n_1}(t) \geq 1/n_3$ for all $t \geq \xi_{(j_N)}$,

$$\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}(t) \leq (1+\varepsilon)^2 t^{1-\delta}.$$

Now, let $B(t) := C_U t^{-b_1}(1-t)^{-b_2}$. Then, by Assumption 5.2.2 and because B is a convex function, we obtain, for all $t \in [\xi_{(j_N)}, \xi_{(n_1)}]$,

$$\begin{aligned} |A_N(t)| &\leq \left[B\left(\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}(t)\right) \vee B\left(\mathbb{G}_{n_1}(t)\right) \right] + B(t) \\ &\lesssim t^{-b_1-\delta}(1-t)^{-b_2-\delta}, \quad \text{a.s.} \end{aligned}$$

Therefore,

$$|A_N(t)| \mathbb{1}\{t \in [\xi_{(j_N)}, \xi_{(n_1)}]\} q(t) \lesssim t^{1/2-b_1-\delta-\delta_1}(1-t)^{1/2-b_2-\delta-\delta_2}.$$

Moreover, by (5.6.24) and Lemma 5.7.2,

$$\int_0^1 t^{1/2-b_1-\delta-\delta_1}(1-t)^{1/2-b_2-\delta-\delta_2}\, dF_Y^{-1}(t) < \infty.$$

Then, by the dominated convergence theorem,

$$\int_{\xi_{(j_N)}}^{\xi_{(n_1)}} |A_N(t)| q(t)\, dF_Y^{-1}(t) \xrightarrow{\text{a.s.}} 0. \tag{5.6.26}$$

Next, since for any $t \in (1/n_3, 1-1/n_1)$,

$$\left| \frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)} \right| = \mathbb{1}\{n_1 \leq n_3\} \left| \frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)} \right| + \mathbb{1}\{n_1 > n_3\} \left| \frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)} \right|,$$

we have

$$\sup_{t\in(1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)}\right| \le 2\sup_{t\in(1/n_3,1-1/n_3)}\left|\frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)}\right| + \mathbb{1}\{n_1>n_3\}\sup_{t\in(1-1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)}\right|$$

By Corollary 4.3.1 and Theorem 3.4 in Csorgo et al. (1986),

$$\sup_{t\in(1/n_3,1-1/n_3)}\left|\frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)}\right| = O_p(1).$$

Also,

$$\begin{aligned}
&\mathbb{1}\{n_1>n_3\}\sup_{t\in(1-1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)}\right| \\
&\quad= \mathbb{1}\{n_1>n_3\}\sup_{t\in(1-1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(\zeta_{(n_3)}-t)}{q(t)}\right| \\
&\quad\le \mathbb{1}\{n_1>n_3\}\sup_{t\in(1-1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(\zeta_{(n_3)}-1)}{q(t)}\right| \\
&\qquad+ \mathbb{1}\{n_1>n_3\}\sup_{t\in(1-1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(1-t)}{q(t)}\right| \\
&\quad\lesssim n_1^{-\delta_2}\left[n_3(\zeta_{(n_3)}-1)+1\right] \\
&\quad= o_p(1).
\end{aligned}$$

Hence,

$$\sup_{t\in(1/n_3,1-1/n_1)}\left|\frac{\sqrt{n_3}(\mathbb{H}_{n_3}^{-1}(t)-t)}{q(t)}\right| = O_p(1).$$

This, together with (5.6.25) and (5.6.26), implies that $R_{1N,2} = o_p(1)$. Now, let us show that $R_{1N,1} = o_p(1)$. Recall that

$$R_{1N,1} := \sqrt{\frac{N}{n_3}}\sqrt{n_3}\int_{\xi_{(1)}}^{\xi_{(j_N)}} A_N(t)\left[\mathbb{H}_{n_3}^{-1}\circ\mathbb{G}_{n_1}(t)-\mathbb{G}_{n_1}(t)\right]dF_Y^{-1}(t),$$

where $j_N = \lceil n_1/n_3\rceil$. If $n_1 \le n_3$, then $j_N = 1$ and $R_{1N,1} = 0$. Thus, assume without loss of generality that $\lambda_3/\lambda_1 > 1$. Whenever $n_1 > n_3$, since $\mathbb{G}_{n_1}$ is monotonically increasing, for all $t \in (\xi_{(1)}, \xi_{(j_N)})$,

$$0 < \frac{1}{n_1} = \mathbb{G}_{n_1}(\xi_{(1)}) \le \mathbb{G}_{n_1}(t) \le \mathbb{G}_{n_1}(\xi_{(j_N)}) \le \frac{1}{n_3}+\frac{1}{n_1} < \frac{2}{n_3}.$$

Therefore, on such interval we have $\mathbb{H}_{n_3}^{-1}\circ\mathbb{G}_{n_1}(t) \in \{\zeta_{(1)}, \zeta_{(2)}\}$. Also, by assumption, there exists $\varepsilon > 0$ such that, for N sufficiently large, $|\frac{n_1}{n_3} - \frac{\lambda_3}{\lambda_1}| < \varepsilon$. By choosing ε sufficiently small, this ensures that for sufficiently large N, $j_N \le \lceil \varepsilon + \lambda_3/\lambda_1\rceil \le$

$\lceil \lambda_3/\lambda_1 \rceil + 1 =: \bar{j}$. Hence, for N sufficiently large,

$$
\begin{aligned}
|R_{1N,1}| &\leq \sqrt{\frac{N}{n_3}}\sqrt{n_3}\mathbb{1}\{n_1 > n_3\} \int_{\xi_{(1)}}^{\xi_{(j_N)}} |A_N(t)| \left|\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}(t) - \mathbb{G}_{n_1}(t)\right| dF_Y^{-1}(t) \\
&\leq \sqrt{\frac{N}{n_3}}\sqrt{n_3}\mathbb{1}\{n_1 > n_3\} \int_{\xi_{(1)}}^{\xi_{(\bar{j})}} |A_N(t)| \left|\mathbb{H}_{n_3}^{-1} \circ \mathbb{G}_{n_1}(t) - \mathbb{G}_{n_1}(t)\right| dF_Y^{-1}(t) \\
&\lesssim \sqrt{\frac{N}{n_3}}\frac{1}{\sqrt{n_3}} \max_{z=1,2} \left\{ n_3 \left|\zeta_{(z)} - \frac{1}{n_1}\right| \vee n_3 \left|\zeta_{(z)} - \frac{\bar{j}}{n_1}\right| \right\} \\
&\qquad \times \left[B(\xi_{(1)}) \vee B(\xi_{(\bar{j})}) \vee B(\zeta_{(1)}) \vee B(\zeta_{(2)}) \vee B(1/n_1) \vee B\left(\bar{j}/n_1\right)\right] \int_{\xi_{(1)}}^{\xi_{(\bar{j})}} dF_Y^{-1}(t),
\end{aligned}
$$

where the third inequality follows by convexity of $u \mapsto |\zeta_{(z)} - u|$, and Assumption 5.2.2. Now, by using the explicit formula for the Mean Absolute Deviation of Beta distributions given in the proof of Lemma 5.7.3, the bound $f(u) \lesssim (C/n_3)u^2$ derived in the same Lemma and applied to $u = 1$ and $u = 2$, and the fact that there exist $c, C > 0$, such that for N sufficiently large $cn_3 \leq n_1 \leq Cn_3$ almost surely, one can show that

$$
\max_{z=1,2} \left\{ n_3 \left|\zeta_{(z)} - \frac{1}{n_1}\right| \vee n_3 \left|\zeta_{(z)} - \frac{\bar{j}}{n_1}\right| \right\} = O_p(1).
$$

Next, by Assumption 5.2.2,

$$
\int_{\xi_{(1)}}^{\xi_{(\bar{j})}} dF_Y^{-1}(t) \lesssim \xi_{(1)}^{-d_1} + \xi_{(\bar{j})}^{-d_1}.
$$

By Assumption 5.2.2 again, there exists $\delta > 0$ such that $1/2 > \delta + b_1 + d_1$. Hence,

$$
\begin{aligned}
&n_3^{-1/2} \left[B(\xi_{(1)}) \vee B(\xi_{(\bar{j})}) \vee B(\zeta_{(1)}) \vee B(\zeta_{(2)}) \vee B(1/n_1) \vee B\left(\bar{j}/n_1\right)\right] \int_{\xi_{(1)}}^{\xi_{(\bar{j})}} dF_Y^{-1}(t) \\
&\lesssim \frac{1}{n_3^{\delta}} \Big[(n_3\xi_{(1)})^{-b_1-d_1} + (n_3\xi_{(\bar{j})})^{-b_1-d_1} + (n_3\xi_{(1)})^{-b_1}(n_3\xi_{(\bar{j})})^{-d_1} + (n_3\xi_{(1)})^{-d_1}(n_3\xi_{(\bar{j})})^{-b_1} \\
&\quad + (n_3\zeta_{(1)})^{-b_1}(n_3\xi_{(1)})^{-d_1} + (n_3\zeta_{(2)})^{-b_1}(n_3\xi_{(\bar{j})})^{-d_1} + (n_3\zeta_{(1)})^{-b_1}(n_3\xi_{(\bar{j})})^{-d_1} \\
&\quad + (n_3\zeta_{(2)})^{-b_1}(n_3\xi_{(1)})^{-d_1} + (n_3/n_1)^{-b_1}(n_3\xi_{(1)})^{-d_1} + (n_3\bar{j}/n_1)^{-b_1}(n_3\xi_{(\bar{j})})^{-d_1} \\
&\quad + (n_3/n_1)^{-b_1}(n_3\xi_{(\bar{j})})^{-d_1} + (n_3\bar{j}/n_1)^{-b_1}(n_3\xi_{(1)})^{-d_1} \Big] \\
&\lesssim \frac{O_p(1)}{n_3^{\delta}}.
\end{aligned}
$$

The last inequality follows from $n_3/n_1 \xrightarrow{p} \lambda_3/\lambda_1$ and the well-known result (see, e.g. the proof of the DKW inequality in Donsker, 1952) that if $(W_i)_{i=1,\ldots,n} \overset{iid}{\sim} \mathcal{U}([0,1])$, then for any $k \in \{1,\ldots,n\}$,

$$
W_{(k)} \overset{d}{=} \frac{S_k}{S_{n+1}},
$$

where $S_j := T_1 + \ldots + T_j$, and $T_1, T_2 \ldots$ is an iid sequence of (mean 1) exponentially distributed random variables. This result, combined with the law of large numbers and the continuous mapping theorem, yields that for any $(k,a) \in \{1,\ldots,n\} \times \mathbb{R}$, $(nW_{(k)})^a =$

$O_p(1)$. We conclude that $R_{1N,1} = o_p(1)$.

Fourth step: $R_{2N} = o_p(1)$. Recall that R_{2N} is defined in (5.6.18). We actually prove the stronger result that R_{2N} converges to 0 in L^1. Let $\mathbb{W}_{n_3} = \sqrt{n_3}(\mathbb{H}_{n_3}^{-1} - \mathrm{I})$ and $B_{n_1} = \mathbb{1}\{\xi_{(1)} \leq x < \xi_{(n_1)}\}$. We have, by Fubini-Tonelli's theorem,

$$\begin{aligned}\mathbb{E}[|R_{2N}|] &\leq \sqrt{\frac{N}{n_3}} \int_0^1 \mathbb{E}[|\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x)| \times B_{n_1}] f_U(x) dF_Y^{-1}(x) \\ &\leq \sqrt{\frac{N}{n_3}} \int_0^1 \mathbb{E}[(\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x))^2 \times B_{n_1}]^{1/2} f_U(x) dF_Y^{-1}(x).\end{aligned}$$

We apply the dominated convergence theorem to prove the result. First, note that for all $(x, y) \in (0, 1]^2$,

$$|\mathbb{H}_{n_3}^{-1}(x) - \mathbb{H}_{n_3}^{-1}(y)| \sim \mathrm{Beta}(|\lceil n_3 x \rceil - \lceil n_3 y \rceil|, n_3 - |\lceil n_3 x \rceil - \lceil n_3 y \rceil| + 1),$$

with the convention that the $\mathrm{Beta}(0, n_3 + 1)$ is the Dirac distribution at 0. Hence, for any $k \in \{1, ..., n_1 - 1\}$,

$$\begin{aligned}&\mathbb{E}\left[(\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x))^2 \,|\mathbb{G}_{n_1}(x) = k/n_1\right] \\ &= n_3 \left\{\mathbb{E}\left[\left(\mathbb{H}_{n_3}^{-1}(k/n_1) - \mathbb{H}_{n_3}^{-1}(x) - (k/n_1 - x)\right)^2\right]\right\} \\ &= n_3 \left\{\mathbb{E}\left[\left(\mathbb{H}_{n_3}^{-1}(k/n_1) - \mathbb{H}_{n_3}^{-1}(x) - \frac{1}{n_3+1}\left(\lceil (n_3 k)/n_1 \rceil - \lceil n_3 x \rceil\right)\right.\right.\right. \\ &\quad \left.\left.\left. + \frac{1}{n_3+1}\left(\lceil (n_3 k)/n_1 \rceil - \lceil n_3 x \rceil\right) - (k/n_1 - x)\right)^2\right]\right\} \\ &= n_3 \left\{\mathbb{V}\left[\mathbb{H}_{n_3}^{-1}(k/n_1) - \mathbb{H}_{n_3}^{-1}(x)\right] + \frac{1}{(n_3+1)^2}(\lceil (n_3 k)/n_1 \rceil - (n_3 k)/n_1 + n_3 x - \lceil n_3 x \rceil + x - k/n_1)^2\right\} \\ &= \frac{n_3}{(n_3+1)^2(n_3+2)} \left|\lceil (n_3 k)/n_1 \rceil - \lceil n_3 x \rceil\right| \left(n_3 + 1 - \left|\lceil (n_3 k)/n_1 \rceil - \lceil n_3 x \rceil\right|\right) \quad (5.6.27) \\ &\quad + \frac{n_3}{(n_3+1)^2}(\lceil (n_3 k)/n_1 \rceil - (n_3 k)/n_1 + n_3 x - \lceil n_3 x \rceil + x - k/n_1)^2 \\ &\leq \frac{1}{n_3}\left|\lceil (n_3 k)/n_1 \rceil - \lceil n_3 x \rceil\right| + \frac{2}{n_3}\left[2\left(\lceil (n_3 k)/n_1 \rceil - (n_3 k)/n_1\right)^2 + 2\left(\lceil n_3 x \rceil - n_3 x\right)^2 + \left(\frac{k}{n_1} - x\right)^2\right] \\ &\leq \left|\frac{k}{n_1} - x\right| + \frac{1}{n_3}\left[10 + 2\left(\frac{k}{n_1} - x\right)^2\right], \quad (5.6.28)\end{aligned}$$

where the first inequality follows by convexity and the last by the triangle inequality and because by definition, $|a - \lceil a \rceil| \leq 1$ for all $a \in \mathbb{R}_+$. Now, remark that $B_{n_1} = 1$ iff $n_1 \mathbb{G}_{n_1}(x) \in \{1, \ldots, n_1 - 1\}$. Then, by what precedes,

$$\mathbb{E}\left[(\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x))^2 \times B_{n_1}\right] \leq \mathbb{E}\left[|\mathbb{G}_{n_1}(x) - x|\right] + \frac{1}{n_3}\left[10 + 2\mathbb{V}(\mathbb{G}_{n_1}(x))\right] \quad (5.6.29)$$

$$\to 0.$$

To apply the dominated convergence theorem, we show that there exists $q(.)$ such that for all $n_1 \geq n_0$ and all $x \in [0,1]$,

$$\mathbb{E}[(\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x))^2 \times B_{n_1}]^{1/2} \leq q(x), \tag{5.6.30}$$

with $\int_0^1 q(x) f_U(x) dF_Y^{-1}(x) < \infty$. As above, we focus on a neighborhood of 0. If $x > 1/n_3$, we have, by (5.6.29) and (5.6.11),

$$\mathbb{E}\left[(\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x))^2 \times B_{n_1}\right] \leq 14x.$$

Now suppose that $x < 1/n_3$. Remark that $\mathbb{E}(B_{n_1}) \leq 1 - (1-x)^{n_1} \leq n_1 x$. Then, integrating (5.6.28), we obtain

$$\begin{aligned}\mathbb{E}\left[(\mathbb{W}_{n_3} \circ \mathbb{G}_{n_1}(x) - \mathbb{W}_{n_3}(x))^2 \times B_{n_1}\right] &\leq \mathbb{E}\left[|\mathbb{G}_{n_1}(x) - x|\right] + \frac{1}{n_3}\left[10 n_1 x + 2\mathbb{V}(\mathbb{G}_{n_1}(x))\right] \\ &\leq (\lceil \lambda_3/\lambda_1 \rceil + 1) 14x.\end{aligned}$$

Then we can choose $q(x) = ((\lceil \lambda_3/\lambda_1 \rceil + 1)14x)^{1/2}$ in (5.6.30). By Assumption 5.2.2 and Lemma 5.7.2, we have $\int_0^{1/2} q(x) f_U(x) dF_Y^{-1}(x) < \infty$. The same reasoning applies to the interval $[1/2, 1]$. The result follows.

Fifth step: $R_{3N} = o_p(1)$. Recall that R_{3N} is defined in (5.6.19). Let Λ denote the measure on $(0,1)$ such that $d\Lambda/dF_Y^{-1} = f_U$. Note that $\mathbb{H}_{n_3}(x) \sim$Beta$(\lceil n_3 x \rceil, n_3 + 1 - \lceil n_3 x \rceil)$, thus $\mathbb{E}[\mathbb{H}_{n_3}(x)] = \lceil n_3 x \rceil / (n_3+1)$. Then

$$\mathbb{E}[|R_{3N}|] \leq \sqrt{\frac{N}{n_3}} \int_0^1 [1 - x^{n_1} - (1-x)^{n_1}] \left| \frac{\lceil n_3 x \rceil - (n_3+1)x}{(n_3+1) n_3^{-1/2}} \right| d\Lambda(x).$$

Let $f_N(x)$ denote the integrand. We have $\lim_{N\to\infty} f_N(x) = 0$. Moreover, using $1 - x^{n_1} - (1-x)^{n_1} \leq n_1 x$ and since $n_1 \lesssim n_3$, we obtain, when $x < 1/n_3$,

$$f_N(x) \leq 2 n_3^{1/2} x \leq x^{1/2} \lesssim [x(1-x)]^{1/2}.$$

When $x \in [1/n_3, 1 - 1/n_3]$,

$$f_N(x) \leq \frac{2}{n_3^{1/2}} \lesssim [x(1-x)]^{1/2}.$$

Finally, when $x > 1 - 1/n_3$, then $x > 1 - 1/n_1$, and thus using $1 - x^{n_1} \leq n_1(1-x)$,

$$f_N(x) \leq n_1(1-x) \frac{2}{(n_3+1) n_3^{-1/2}} \leq 2(1-x)^{1/2} \lesssim [x(1-x)]^{1/2}.$$

Moreover, $\int_0^1 [x(1-x)]^{1/2} d\Lambda < \infty$ by Lemma 5.7.2. Thus, by the dominated convergence theorem, $R_{3N} = o_p(1)$.

Sixth step: $R_{4N} = o_p(1)$. Recall that R_{4N} is defined in (5.6.20). We prove the stronger result that R_{4N} converges to 0 in L^1. By Fubini-Tonelli's theorem combined with Assumption 5.2.1 and Jensen's inequality, we have

$$\mathbb{E}[|R_{4N}|] \leq \sqrt{\frac{N}{n_3}} \int_0^1 \sqrt{n_3}\mathbb{E}\left[\left|\mathbb{1}\{x \in [\xi_{(1)}, \xi_{(n_1)}]\} - \mathbb{1}\{x \in [1/n_3, (n_3-1)/n_3]\}\right|\right] \times \mathbb{E}\left[\left(\mathbb{H}_{n_3}^{-1}(x) - \frac{\lceil n_3 x \rceil}{(n_3+1)}\right)^2\right]^{1/2} d\Lambda(x). \tag{5.6.31}$$

Since $\mathbb{H}_{n_3}^{-1}(x) \sim$Beta$(\lceil n_3 x \rceil, n_3 + 1 - \lceil n_3 x \rceil)$, we have

$$\mathbb{E}\left[\left(\mathbb{H}_{n_3}^{-1}(x) - \frac{\lceil n_3 x \rceil}{(n_3+1)}\right)^2\right]^{1/2} = \sqrt{\frac{\lceil n_3 x \rceil (n_3 + 1 - \lceil n_3 x \rceil)}{(n_3+1)^2(n_3+2)}} \lesssim \sqrt{\frac{x(1-x)}{n_3}}.$$

Let $q_N(x)$ denote the first expectation in the integrand. By letting $p_{n_1}(x) := 1 - x^{n_1} - (1-x)^{n_1}$, we have

$$\begin{aligned} q_N(x) &= \mathbb{P}(\xi_{(1)} \leq x \leq \xi_{(n_1)}, x < 1/n_3) + \mathbb{P}(\xi_{(1)} \leq x \leq \xi_{(n_1)}, x > (n_3-1)/n_3) \\ &\quad + \mathbb{P}(\xi_{(1)} > x \cup x > \xi_{(n_1)}, 1/n_3 \leq x \leq (n_3-1)/n_3) \\ &= p_{n_1}(x)\left[\mathbb{1}\{x < 1/n_3\} + \mathbb{1}\{1-x < 1/n_3\}\right] + (1 - p_{n_1}(x))\mathbb{1}\{1/n_3 \leq x \leq (n_3-1)/n_3\}. \end{aligned}$$

Let $f_N(x)$ denote the integrand in the right-hand side of (5.6.31). For all $x \in (0,1)$, $\lim_{N\to\infty} p_{n_1}(x) = 1$ so from what precedes, $\lim_{N\to\infty} f_N(x) = 0$ for all $x \in [0,1]$. Moreover, using $q_N(x) \leq 1$, we get

$$f_N(x) \lesssim [x(1-x)]^{1/2},$$

with $\int_0^1 [x(1-x)]^{1/2} d\Lambda < \infty$ by Lemma 5.7.2. The result follows by the dominated convergence theorem.

Seventh step: asymptotic normality of J_{2n_3}. Let $I_{in_3} = [(i-1)/n_3, i/n_3)$. First, note that

$$-\sqrt{n_3} J_{2n_3} = \sum_{i=1}^{n_3} a_{in_3} \left(\zeta_{(i)} - \frac{i}{(n_3+1)}\right), \tag{5.6.32}$$

where $a_{1n_3} = a_{n_3 n_3} = 0$, and, for all $i \in \{2, \ldots, n_3 - 1\}$, $a_{in_3} = \sqrt{n_3}\Lambda(I_{in_3})$. We now verify that the necessary and sufficient conditions given by Hecker (1976) for the asymptotic normality of the $L-$statistic in (5.6.32) hold in our case. Let us define

$$\sigma_{n_3}^2 = \frac{1}{n_3+2} \sum_{j=1}^{n_3} \sum_{k=1}^{n_3} a_{jn_3} a_{kn_3} \left[\left(\frac{j}{n_3+1} \wedge \frac{k}{n_3+1}\right) - \frac{jk}{(n_3+1)^2}\right].$$

We have to prove that

$$\lim_{n_3 \to \infty} \frac{\max_{1 \leq i \leq n_3} \left|\sum_{j=i}^{n_3} a_{jn_3}\right|}{n_3 \sigma_{n_3}} = 0. \tag{5.6.33}$$

First, by Assumption 5.2.2 and Lemma 5.7.2, there exists $\delta < 1/2$ such that

$$\int_0^1 t^{\delta-b_1}(1-t)^{\delta-b_2}\, dF_Y^{-1}(t) < +\infty.$$

Now, because $a_{in_3} \geq 0$, we have, for all $n_3 \geq 2$,

$$\begin{aligned}
\max_{1\leq i\leq n_3}\left|\sum_{j=i}^{n_3} a_{jn_3}\right| &= \sqrt{n}_3 \sum_{j=2}^{n_3-1} \Lambda\left(I_{jn_3}\right) \\
&= \sqrt{n}_3 \int_{1/n_3}^{(n_3-1)/n_3} f_U(t)\, dF_Y^{-1}(t) \\
&\leq C_U\sqrt{n_3}\int_{1/n_3}^{(n_3-1)/n_3} t^{-b_1}(1-t)^{-b_2}\, dF_Y^{-1}(t) \\
&\leq C_U 2^\delta n_3^{1/2+\delta}\int_{1/n_3}^{(n_3-1)/n_3} t^{\delta-b_1}(1-t)^{\delta-b_2}\, dF_Y^{-1}(t) \\
&\leq C_U 2^\delta n_3^{1/2+\delta}\int_0^1 t^{\delta-b_1}(1-t)^{\delta-b_2}\, dF_Y^{-1}(t),
\end{aligned}$$

where the first inequality follows by Assumption 5.2.2 and the second uses the fact that $[t(1-t)]^\delta \geq 1/(2n_3)^\delta$ for all $t \in [1/n_3, 1-1/n_3]$. Therefore,

$$\max_{1\leq i\leq n}\left|\sum_{j=i}^{n_3} a_{jn_3}\right| = O(n_3^{1/2+\delta}). \tag{5.6.34}$$

Next, we have

$$\begin{aligned}
\sigma_{n_3}^2 &= \frac{n_3}{n_3+2}\sum_{j=2}^{n_3-1}\sum_{k=2}^{n_3-1}\Lambda\left(I_{jn_3}\right)\Lambda\left(I_{kn_3}\right)\left(\frac{j}{n_3+1}\wedge\frac{k}{n_3+1}-\frac{jk}{(n_3+1)^2}\right) \\
&= \frac{n}{n_3+2}\int_0^1\int_0^1 f_{n_3}(x,y)d\Lambda(x)d\Lambda(y),
\end{aligned}$$

where $f_{n_3}(x,y) = \frac{j}{n_3+1}\wedge\frac{k}{n_3+1}-\frac{jk}{(n_3+1)^2}$ when $(x,y) \in I_{jn_3}\times I_{kn_3}$, $1 < j\wedge k \leq j\vee k < n_3$, $f_{n_3}(x,y) = 0$ otherwise. For any $(x,y) \in (0,1)^2$, $f_{n_3}(x,y) \to f(x,y) := x\wedge y - xy$. Moreover, for any $(x,y) \in I_{jn_3}\times I_{kn_3}$, $1 < j\wedge k \leq j\vee k < n_3$,

$$\begin{aligned}
\frac{j}{n_3+1}\wedge\frac{k}{n_3+1} &\leq 2(x\wedge y), \\
1-\frac{j}{n_3+1}\vee\frac{k}{n_3+1} &\leq 2\left(1-x\vee y\right).
\end{aligned}$$

Thus, $f_{n_3}(x,y) \leq 4f(x,y)$ for all $(x,y) \in [1/n_3, 1-1/n_3]^2$. This inequality also holds for $(x,y) \in [0,1]^2\backslash[1/n_3, 1-1/n_3]^2$ since $f_{n_3}(x,y) = 0$ for such (x,y). Because $x\wedge y \leq (xy)^{1/2}$ and $1-x\vee y \leq [(1-x)(1-y)]^{1/2}$, we have $f(x,y) \leq [x(1-x)y(1-y)]^{1/2}$. Moreover, by Lemma 5.7.2, $\int_0^1[\mathrm{I}(1-\mathrm{I})]^{1/2}d\Lambda < \infty$. Thus, by the dominated

convergence theorem,

$$\lim_{n_3\to\infty}\sigma_{n_3}^2=\sigma^2:=\int_0^1\int_0^1(x\wedge y-xy)d\Lambda(x)d\Lambda(y)>0. \tag{5.6.35}$$

Combined with (5.6.34), this implies (5.6.33). Thus, by Theorem 1 of Hecker (1976) and (5.6.35) again,

$$-\sqrt{n_3}J_{2n_3}\xrightarrow{d}\mathcal{N}(0,\sigma^2).$$

Eighth step: conclusion. By the previous steps and the proof of Theorem 5.2.2, we have

$$\sqrt{N}\left(\check{\theta}_{n_1,n_2,n_3}-\theta_0\right)=\sqrt{\frac{\lambda_1}{n_1}}\sum_{i=1}^{n_1}\eta_i+\sqrt{\frac{\lambda_2}{n_2}}\sum_{i=1}^{n_2}\varepsilon_i+\sqrt{\lambda_3 n_3}J_{2n_3}+o_p(1).$$

As shown in the proof of Theorem 5.2.2, the first term on the right-hand side is asymptotically normal. The second term is also asymptotically normal by the previous step. Moreover, by Assumption 5.3.1, J_{2n_3} is independent of the $(\eta_i,\varepsilon_i)_{i\geq 1}$. Therefore, the vector $\left(\sum_{i=1}^{n_1}(\eta_i)/\sqrt{n_1}+\sum_{i=1}^{n_2}(\varepsilon_i)/\sqrt{n_2},\sqrt{n_3}J_{2n_3}\right)$ converges jointly in distribution to two independent normal variables distributions. The result follows. □

5.7 Technical Lemmas

In Theorems 1 and 2, we use the following lemma, which is established in Proposition 1 of Falkner and Teschl (2012).

Lemma 5.7.1 *Let g be some Borel measurable function on $[0,1]$, and F,Q be cdf's on $[0,1]$. Then, for any $0\leq a\leq b\leq 1$,*

$$\int_{Q(a)}^{Q(b)}g\circ Q^{-1}\,dF=\int_a^b g\,dF\circ Q. \tag{5.7.1}$$

Lemma 5.7.2 *Suppose that Assumption 5.2.2 holds and that $a_1>d_1$ and $a_2>d_2$, then $\int_0^1 x^{a_1}(1-x)^{a_2}\,dF_Y^{-1}(x)<\infty$.*

Proof: first, we have

$$\int_0^1 x^{a_1}(1-x)^{a_2}\,dF_Y^{-1}(x)=\int_{\mathbb{R}}F_Y(u)^{a_1}(1-F_Y(u))^{a_2}\,du.$$

By Assumption 5.2.2 (ii), for all $u\in\mathbb{R}$:

$$|u|\leq CF_Y(u)^{-d_1}(1-F_Y(u))^{-d_2}.$$

Fix $\varepsilon>0$. Then, for all $u\leq -1\wedge F_Y^{-1}(\varepsilon)$, $F_Y(u)\leq C^{1/d_1}(1-\varepsilon)^{-d_2/d_1}|u|^{-1/d_1}$. Thus:

$$\int_{-\infty}^{-1\wedge F_Y^{-1}(\varepsilon)}F_Y(u)^{a_1}(1-F_Y(u))^{a_2}\,du\leq C^{1/d_1}(1-\varepsilon)^{-d_2/d_1}\int_{-\infty}^{-1\wedge F_Y^{-1}(\varepsilon)}|u|^{-a_1/d_1}\,du<\infty,$$

since $d_1 < a_1$. A similar reasoning shows that $\int_{1\vee F_Y^{-1}(1-\varepsilon)}^{\infty} F_Y(u)^{a_1}(1-F_Y(u))^{a_2}\,du < \infty$, using $d_2 < a_2$. □

We recall that $\overline{\mathbb{G}}_{n_1}$ is defined as $\overline{\mathbb{G}}_{n_1}(x) = \mathbb{G}_{n_1}(x) + \sum_{i=1}^{n_1-1} \mathbb{1}\{\xi_{(i)} < x < \xi_{(i+1)}\}/n_1$.

Lemma 5.7.3 *(Useful properties of $\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}$) There exists $\delta \in (0, 1/2)$ and $n_0 \in \mathbb{N}$ such that for all $0 < x < \delta$ and all $n \geq n_0$,*

$$\mathbb{E}\left[\left|\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) - x\right|\right] \lesssim x. \tag{5.7.2}$$

Moreover, for any $\eta > 0$, there exists n_0' such that for all $n \geq n_0'$ and for all $0 < x < \delta$,

$$\mathbb{P}(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) > 1/2) \lesssim x^{1-\eta}. \tag{5.7.3}$$

Inequalities (5.7.2)-(5.7.3) *hold if we replace x by $1-x$, using possibly another δ and n_0.*

Proof: Let us define

$$\widetilde{\mathbb{G}}_{n_1}(x) = \mathbb{G}_{n_1}(x) + \frac{\mathbb{1}\{0 < \mathbb{G}_{n_1}(x) < 1\}}{n}. \tag{5.7.4}$$

Observe that for a given $x \in [0,1]$, we have, with probability one, $\overline{\mathbb{G}}_{n_1}(x) = \widetilde{\mathbb{G}}_{n_1}(x)$. Then, recalling that $p_{n_1}(x) = [1 - x^{n_1} - (1-x)^{n_1}]/n_1$,

$$\mathbb{E}[\overline{\mathbb{G}}_{n_1}(x)] = x + p_{n_1}(x).$$

By the triangle inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) - x\right|\right] \\
&\leq \mathbb{E}\left[\left|\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) - \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil}{n_3+1}\right|\right] + \mathbb{E}\left[\left|\frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil}{n_3+1} - x\right|\right].
\end{aligned} \tag{5.7.5}$$

Consider the second term first. Suppose first that $n_3 x \leq 1$. Let $\lambda = n_3/n_1$ and $B \sim$Binomial(n_1, x). We have

$$\mathbb{E}\left[\left|\frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil}{n_3+1} - x\right|\right] \leq \frac{\mathbb{E}\left[|\lceil \lambda(B + \mathbb{1}\{n_1 > B > 0\}\rceil - \lambda n_1 x|\right] + x}{n_3+1}. \tag{5.7.6}$$

Next, using $\lambda n_1 x \leq 1$, $\lceil \lambda(k+1) \rceil \leq \lambda(k+1)+1$ and $\mathbb{P}(B>0) \leq n_1 x$, we get

$$\begin{aligned}
&\mathbb{E}[|\lceil \lambda(B+\mathbb{1}\{n_1 > B > 0\}) \rceil - \lambda n_1 x|] \\
&\leq \lambda n_1 x \mathbb{P}(B=0) + \sum_{k=1}^{n_1} \left(\lceil \lambda(k+1) \rceil - \lambda n_1 x \right) \mathbb{P}(B=k) \\
&\leq \lambda n_1 x (1 - \mathbb{P}(B>0)) + \sum_{k=1}^{n_1} (\lambda(k+1) + 1 - n_1 x) \mathbb{P}(B=k) \\
&\leq n_1 x (2\lambda - (\lambda+1)\mathbb{P}(B>0)) + (\lambda+1)\mathbb{P}(B>0) \\
&\leq (3\lambda+1) n_1 x. \qquad (5.7.7)
\end{aligned}$$

Combining (5.7.6), (5.7.7) and $n_1 \lesssim n_3$ yields

$$\mathbb{E}\left[\left| \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil}{n_3+1} - x \right| \right] \lesssim x.$$

Now, suppose that $n_3 x > 1$. We have

$$\begin{aligned}
\mathbb{E}\left[\left| \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil}{n_3+1} - x \right| \right] &\leq \frac{1}{n_3+1} + \frac{n_3}{n_3+1} \mathbb{E}\left[|\mathbb{G}_{n_1}(x) - x| \right] + \frac{n_3 p_{n_1}(x)}{n_3+1} + \frac{x}{n_3+1} \\
&\leq \frac{1}{n_3+1} + \mathbb{E}\left[|\mathbb{G}_{n_1}(x) - x| \right] + 2x,
\end{aligned}$$

where the first inequality uses the triangle inequality, $|\lceil a \rceil - a| \leq 1$ for all $a \in \mathbb{R}_+$, and $\overline{\mathbb{G}}_{n_1}(x) = \widetilde{\mathbb{G}}_{n_1}(x)$ with probability one, and the second inequality follows by $p_{n_1}(x) \leq x$. Then, using $n_3 + 1 > 1/x$ and (5.6.11), which holds for all $x \in (0, \widetilde{\delta})$, we also obtain in this case

$$\mathbb{E}\left[\left| \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil}{n_3+1} - x \right| \right] \lesssim x. \qquad (5.7.8)$$

Now, let us bound the first term of (5.7.5). Again, because $\mathbb{H}_{n_3}^{-1}(x) \sim \text{Beta}(\lceil n_3 x \rceil, n_3 + 1 - \lceil n_3 x \rceil)$, we have

$$\mathbb{E}\left[\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) \,|\overline{\mathbb{G}}_{n_1}(x) \right] = \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil}{n_3+1}.$$

Moreover, any $Z \sim \text{Beta}(a,b)$ satisfies $\mathbb{E}[|Z - \mathbb{E}(Z)|] = 2a^a b^b / (B(a,b)(a+b)^{a+b+1})$. Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left| \mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) - \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil}{n_3+1} \right| \middle| \overline{\mathbb{G}}_{n_1}(x) \right] \\
&= \frac{2 \lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil^{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil} (n_3 + 1 - \lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil)^{n_3 + 1 - \lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil}}{B(\lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil, n_3 + 1 - \lceil n_3 \overline{\mathbb{G}}_{n_1}(x) \rceil)(n_3+1)^{n_3+2}} \mathbb{1}\{\overline{\mathbb{G}}_{n_1}(x) > 0\}. \qquad (5.7.9)
\end{aligned}$$

Now, let $f(u) = 2u^u(n_3+1-u)^{n_3+1-u}/[B(u, n_3+1-u)(n_3+1)^{n_3+2}]$ for $u \in [0, n_3]$. Then:

$$\mathbb{E}\left[\left|\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) - \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil}{n_3+1}\right|\right] = \mathbb{E}\left[f\left(\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil\right)|\overline{\mathbb{G}}_{n_1}(x) > 0\right]\mathbb{P}(\overline{\mathbb{G}}_{n_1}(x) > 0).$$

Now, Stirling's formula gives the following bound for the beta function $B(\cdot,\cdot)$ (see, e.g., pp. 263, Ex. 45, Whittaker and Watson, 1996):

$$\frac{1}{B(x,y)} < \frac{1}{\sqrt{2\pi}}\frac{(x+y)^{x+y-1/2}}{x^{x-1/2}y^{y-1/2}}, \quad \forall x, y > 0. \tag{5.7.10}$$

Plugging (5.7.10) for $x = u$ and $y = n_3+1-u$ in the definition of $f(u)$, we have for all $1 \le u \le n_3$

$$\begin{aligned} f(u) &\le \frac{1}{\sqrt{2\pi}}\frac{2u^u(n_3+1-u)^{n_3+1-u}(n_3+1)^{n_3+1/2}}{u^{u-1/2}(n_3+1-u)^{n_3+1/2-u}(n_3+1)^{n_3+2}} \\ &\lesssim \frac{u^{1/2}(n_3+1-u)^{1/2}}{(n_3+1)^{3/2}} \\ &\lesssim \frac{u}{n_3}, \end{aligned}$$

where the last inequality uses $u^{1/2} \le u$ and $(n_3+1-u) \le n_3+1$ for all $1 \le u \le n_3$. Hence,

$$\begin{aligned} \mathbb{E}\left[\left|\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) - \frac{\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil}{n_3+1}\right|\right] &\lesssim \frac{1}{n_3}\mathbb{E}(\lceil n_3 \overline{\mathbb{G}}_{n_1}(x)\rceil|\overline{\mathbb{G}}_{n_1}(x) > 0)\mathbb{P}(\overline{\mathbb{G}}_{n_1}(x) > 0) \\ &\lesssim \frac{1}{n_3}\left(1 + n_3\mathbb{E}(\overline{\mathbb{G}}_{n_1}(x)|\overline{\mathbb{G}}_{n_1}(x) > 0)\right)\mathbb{P}(\overline{\mathbb{G}}_{n_1}(x) > 0) \\ &\lesssim \frac{1}{n_3}(n_1 x + n_3(x + p_{n_1}(x))) \\ &\lesssim x. \end{aligned}$$

where we have used $\lceil a\rceil \le a+1$, $\mathbb{P}(\overline{\mathbb{G}}_{n_1}(x) > 0) \le n_1 x$ and $\mathbb{P}(\overline{\mathbb{G}}_{n_1}(x) > 0)\mathbb{E}(\overline{\mathbb{G}}_{n_1}(x|\overline{\mathbb{G}}_{n_1}(x) > 0) = \mathbb{E}(\overline{\mathbb{G}}_{n_1}(x))$.

We now turn to Equation (5.7.3). Suppose that there exists a constant $\underline{c} > 0$ such that $\underline{c} \le n_3/n_1$ (or simply assume that $\lambda_1/\lambda_3 > \underline{c}$). Because $|\overline{\mathbb{G}}_{n_1}(x) - \mathbb{G}_{n_1}(x)| \le 1/n_1$, $\overline{\mathbb{G}}_{n_1}(x) > \mathbb{H}_{n_3}(1/2)$ implies $\mathbb{G}_{n_1}(x) > \mathbb{H}_{n_3}(1/2) - 1/n_1$. Moreover, $\mathbb{H}_{n_3}^{-1}(a) < b$ iff

$a < \mathbb{H}_{n_3}(b)$. Then, by Kiefer's and Hoeffding's inequalities,

$$\begin{aligned}
\mathbb{P}(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) > 1/2) &= \mathbb{E}\left[\mathbb{P}(\overline{\mathbb{G}}_{n_1}(x) > \mathbb{H}_{n_3}(1/2) | \mathbb{H}_{n_3}(1/2))\right] \\
&\leq \mathbb{E}\left[\mathbb{P}(\mathbb{G}_{n_1}(x) \geq \mathbb{H}_{n_3}(1/2) - 1/n_1 | \mathbb{H}_{n_3}(1/2))\right] \\
&\leq \mathbb{E}\left[(xe)^{n_1(\mathbb{H}_{n_3}(1/2) - 1/n_1 - x)^2}\right] \\
&\leq xe + \mathbb{P}\left(\mathbb{H}_{n_3}(1/2) - x < 1/\sqrt{n_1}\right) \\
&\leq xe + \exp\left(-2(\sqrt{n_3}(x - 1/2) + \sqrt{n_3/n_1})^2\right) \\
&= xe + \exp\left(-2n_3(x - 1/2 + 1/\sqrt{n_1})^2\right).
\end{aligned}$$

Let $\overline{\delta} \in (0, e^{-1}]$ and fix $\delta = \overline{\delta}/2$ and $n_0 \geq (2/\overline{\delta})^2$. Then, for all $n_1 \geq n_0$ and any $0 < x \leq \delta$, we have

$$\begin{aligned}
\left|x - 1/2 + \frac{1}{\sqrt{n_1}}\right| &= \frac{1}{2} - (x + 1/\sqrt{n_1}) \\
&\geq \frac{1}{2} - \overline{\delta}.
\end{aligned}$$

Let $C = 2\underline{c}(1/2 - \overline{\delta})^2$ and suppose first that $x \geq \exp(A - Cn_1)$ for some A. Then some algebra shows that, for N sufficiently large,

$$\mathbb{P}(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) > 1/2) \lesssim x.$$

Now assume that $x < \exp(A - Cn_1)$. Then,

$$\begin{aligned}
\mathbb{P}(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) > 1/2) \leq& \mathbb{P}(\mathbb{G}_{n_1}(x) \geq 1/n_1) \\
=& 1 - (1 - x)^{n_1} \\
\leq& n_1 x \\
\leq& \frac{A - \ln x}{C} x.
\end{aligned}$$

For any $\eta > 0$, we have $-\ln x \lesssim x^{-\eta}$. Thus, $\mathbb{P}(\mathbb{H}_{n_3}^{-1} \circ \overline{\mathbb{G}}_{n_1}(x) > 1/2) \lesssim x^{1-\eta}$ □

Lemma 5.7.4 *(Bounds on moments involving F_U) Suppose that Assumption 5.2.2 holds and a random variable $Q_n(x)$ satisfies, for some $0 < \delta < 1/2$ and all $0 < x < \delta$, $\mathbb{E}[|Q_n(x) - x|] \lesssim x$ and $\mathbb{P}(Q_n(x) > 1/2) \lesssim x^{1-b_1}$. Then, for such $x \in (0, \delta)$, $\mathbb{E}[|F_U(Q_n(x)) - F_U(x)|] \lesssim x^{1-b_1}$. The latter inequality holds if we replace x by $1 - x$, using possibly another δ.*

Proof of Lemma 5.7.4: first, remark that for $x < 1/2$, $F_U(x) \lesssim x^{1-b_1}$. Then,

$$\begin{aligned}\mathbb{E}[|F_U(x) - F_U(Q_n(x))|] \leq & \mathbb{E}[\mathbb{1}\{x > Q_n(x)\}|F_U(x) - F_U(Q_n(x))|] + \mathbb{P}(Q_n(x) > 1/2) \\ & + \mathbb{E}\left[\mathbb{1}\{Q_n(x) \in [x, 1/2]\}|F_U(x) - F_U(Q_n(x))|\right] \\ \lesssim & F_U(x) + x^{1-b_1} + \mathbb{E}\left[\mathbb{1}\{Q_n(x) \in [x, 1/2]\}|F_U(x) - F_U(Q_n(x))|\right] \\ \lesssim & x^{1-b_1} + \mathbb{E}\left[\mathbb{1}\{Q_n(x) \in [x, 1/2]\}|F_U(x) - F_U(Q_n(x))|\right].\end{aligned}$$

Now, if $Q_n(x) \in [x, 1/2)$, by the mean value theorem, there exists $X_n \in (x, 1/2)$ such that

$$F_U(x) - F_U(Q_n(x)) = f_U(X_n)(x - Q_n(x)).$$

Moreover, by Assumption 5.2.2 and $x < \delta$, $f_U(X_n) \lesssim x^{-b_1}$. Then, using $\mathbb{E}[|Q_n(x) - x|] \lesssim x$,

$$\mathbb{E}\left[\mathbb{1}\{Q_n(x) \in [x, 1/2]\}|F_U(x) - F_U(Q_n(x))|\right] \lesssim x^{1-b_1}.$$

The result follows. □

Bibliography

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Aghion, P., N. Bloom, R. Blundell, R. Griffith, and P. Howitt (2005, 05). 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. Technical report.

Angrist, J. and J. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D. and G. W. Imbens (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association 90*(430), 431–442.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies 58*(2), 277–297.

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.

Armstrong, T. B., M. Weidner, and A. Zeleneev (2022). Robust estimation and inference in panels with interactive fixed effects.

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.

Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica 90*(1), 347–365.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. (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.

Bai, J. and S. Ng (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics 212*(1), 78–96. Big Data in Dynamic Predictive Econometric Modeling.

Balakrishnan, N. and M. Leung (1988). Order statistics from the type i generalized logistic distribution. *Communications in Statistics-Simulation and Computation 17*(1), 25–50.

Beck, A. and L. Tetruashvili (2013). On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization 23*(4), 2037–2060.

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political Economy 81*(4), 813–846.

Becker, G. S. (1974). A theory of marriage: Part ii. *Journal of Political Economy 82*(2), S11–S26.

Berend, D. and A. Kontorovich (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters 83*(4), 1254–1259.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fenmlm. Technical report.

Berry, S., A. Gandhi, and P. Haile (2013). Connected substitutes and invertibility of demand. *Econometrica 81*(5), 2087–2111.

Bertsekas, D. (2016). *Nonlinear Programming* (3^{rd} ed.). Athena scientific optimization and computation series. Athena Scientific.

Bester, C. A. and C. B. Hansen (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics 190*(1), 197–208.

Beyhum, J. and E. Gautier (2019). Square-root nuclear norm penalized estimator for panel data models with approximately low-rank unobserved heterogeneity.

Beyhum, J. and E. Gautier (2023). Factor and factor loading augmented estimators for panel regression with possibly nonstrong factors. *Journal of Business & Economic Statistics 41*(1), 270–281.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705 – 1732.

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica 58*(6), 1443–1458.

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

Boneva, L. and O. Linton (2017). A discrete-choice model for large heterogeneous panels with interactive fixed effects with an application to the determinants of corporate bond issuance. *Journal of Applied Econometrics 32*(7), 1226–1243.

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. (2012). Functional differencing. *Econometrica 80*(4), 1337–1385.

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 (2017a, June). Binarization for panel models with fixed effects. CeMMAP working papers CWP31/17.

Botosaru, I. and C. Muris (2017b, June). Binarization for panel models with fixed effects. CeMMAP working papers CWP31/17.

Brender, A. and A. Drazen (2008). How do budget deficits and economic growth affect reelection prospects? evidence from a large panel of countries. *The American Economic Review 98*(5), 2203–2220.

Brownlees, C., G. S. Guðmundsson, and G. Lugosi (2022). Community detection in partial correlation network models. *Journal of Business & Economic Statistics 40*(1), 216–226.

Bryant, P. and J. A. Williamson (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika 65*(2), 273–281.

Candelaria, L. E. (2020). A semiparametric network formation model with unobserved linear heterogeneity.

Carrère, C., M. Mrázová, and J. P. Neary (2020). Gravity without apology: the science of elasticities, distance and trade. *The Economic Journal 130*(628), 880–910.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies 47*(1), 225–238.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–304.

Chamberlain, G. (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. (2016). Estimation of nonlinear panel models with multiple unobserved effects. Technical report.

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 (2021). Payment instrument choice with scanner data: An mmalgorithm for fixed effects in multinomial logit models. Technical report.

Chen, N. and D. Novy (2021). Gravity and heterogeneous trade cost elasticities.

Chen, X., Y. Fan, and V. Tsyrennikov (2006). Efficient estimation of semiparametric multivariate copula models. *Journal of the American Statistical Association 101*(475), 1228–1240.

Cheng, X., F. Schorfheide, and P. Shao (2021). Clustering for multi-dimensional heterogeneity. Technical report.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013, 3). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Chernozhukov, V., C. Hansen, Y. Liao, and Y. Zhu (2019). Inference for heterogeneous effects using low-rank estimation of factor slopes.

Chetverikov, D. and E. Manresa (2021). Spectral and post-spectral estimators for grouped panel data models. Technical report.

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

Correia, S., P. Guimarães, and T. Zylkin (2020). Fast poisson estimation with high-dimensional fixed effects. *The Stata Journal 20*(1), 95–115.

Csorgo, M., S. Csorgo, L. Horváth, and D. M. Mason (1986). Weighted empirical and quantile processes. *The Annals of Probability*, 31–85.

Czarnowske, D. and A. Stammann (2020). Fixed effects binary choice models: Estimation and inference with long panels.

Davezies, L., X. D'Haultfoeuille, and L. Laage (2022). Identification and estimation of average marginal effects in fixed effects logit models.

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, September). 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. *Annual Review of Economics 12*(1), 775–799.

Deb, P. and P. Trivedi (1997, 05). 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*.

D'Haultfoeuille, X. and A. Iaria (2016). A convenient method for the estimation of the multinomial logit model with fixed effects. *Economics Letters 141*, 77–79.

D'Haultfœuille, X., A. Wang, P. Février, and L. Wilner (2022). Estimating the gains (and losses) of revenue management. *arXiv preprint arXiv:2206.04424*.

Dominguez, M. A. and I. N. Lobato (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica 72*(5), 1601–1615.

Donsker, M. D. (1952). Justification and Extension of Doob's Heuristic Approach to the Kolmogorov- Smirnov Theorems. *The Annals of Mathematical Statistics 23*(2), 277 – 281.

Dubois, P., R. Griffith, and M. O'Connell (2020). How well targeted are soda taxes? *American Economic Review 110*(11), 3661–3704.

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

Evdokimov, K. (2010). Identification and estimation of a nonparametric panel data model with unobserved heterogeneity.

Evdokimov, K. (2011). Nonparametric identification of a nonlinear panel model with application to duration analysis with multiple spells.

Falkner, N. and G. Teschl (2012). On the substitution rule for lebesgue–stieltjes integrals. *Expositiones Mathematicae 30*(4), 412 – 418.

Fernandez-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics 150*(1), 71 – 85.

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.

Fernández-Val, I. and M. Weidner (2018, August). 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.

Gale, D. and H. Nikaido (1965). The jacobian matrix and global univalence of mappings. *Mathematische Annalen 159*(2), 81–93.

Gallant, A. R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the econometric society*, 363–390.

Gao, J., F. Liu, and B. Peng (2020). Binary response models for heterogeneous panel data with interactive fixed effects. *arXiv preprint arXiv:2012.03182*.

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. (2020). Nonparametric identification in index models of link formation. *Journal of Econometrics 215*(2), 399–413.

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

Gaure, S. (2013). OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis 66*, 8 – 18. Description of the projection methods used in 'lfe'.

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*. Cambridge, MA, USA: MIT Press.

Gourieroux, C. and A. Monfort (1995). *Statistics and Econometric Models*. Cambridge University Press.

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, November). Product market competition, creative destruction and innovation. LSE Research Online Documents on Economics 113816.

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.

Guimaraes, P. and P. Portugal (2010). A simple feasible procedure to fit models with high-dimensional fixed effects. *The Stata Journal 10*(4), 628–649.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica 11*(1), 1–12.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica 12*, iii–115.

Hahn, J. (1997). A note on the efficient semiparametric estimation of some exponential panel models. *Econometric Theory 13*(4), 583–588.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and t are large. *Econometrica 70*(4), 1639–1657.

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.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hashmi, A. (2013). Competition and innovation: The inverted-u relationship revisited. *The Review of Economics and Statistics 95*(5), 1653–1668.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

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.

Hecker, H. (1976, 11). A characterization of the asymptotic normality of linear combinations of order statistics from the uniform distribution. *Ann. Statist. 4*(6), 1244–1246.

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.

Heckman, J. J. (1981, July-Dece). Heterogeneity and State Dependence. In *Studies in Labor Markets*, NBER Chapters, pp. 91–140. National Bureau of Economic Research, Inc.

Helpman, E., M. Melitz, and Y. Rubinstein (2008). Estimating trade flows: Trading partners and trading volumes. *The quarterly journal of economics 123*(2), 441–487.

Hicks, J. (1939). *Value and Capital: An Inquiry Into Some Fundamental Principles of Economic Theory*. Clarendon Press.

Higgins, A. (2022). Fixed t estimation of linear panel data models with interactive fixed effects.

Hinz, J., A. Hudlet, and J. Wanner (2019). Separating the wheat from the chaff: Fast estimation of glms with high-dimensional fixed effects. *Unpublished Working Paper*.

Hoderlein, S. and H. White (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics 168*(2), 300–314.

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

Hong, Y. and H. White (1995). Consistent specification testing via nonparametric series regression. *Econometrica 63*(5), 1133–1159.

Honoré, B. E. and J. L. Powell (1994). Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics 64*(1), 241–278.

Honoré, B. E. and M. Weidner (2020). Moment conditions for dynamic panel logit models with fixed effects. arXiv eprint 2005.05942.

Honore, B. E. and A. Lewbel (2002). Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica 70*(5), 2053–2063.

Hospido, L. (2012). Estimating nonlinear models with multiple fixed effects: A computational note*. *Oxford Bulletin of Economics and Statistics 74*(5), 760–775.

Hsiao, C. (2014). *Analysis of Panel Data* (3 ed.). Econometric Society Monographs. Cambridge University Press.

Hsu, S.-H. and C.-M. Kuan (2011). Estimation of conditional moment restrictions without assuming parameter identifiability in the implied unconditional moments. *Journal of Econometrics 165*(1), 87 – 99.

Iaria, A. and A. Wang (2021). An empirical model of quantity discounts with large choice sets. *Working Paper*.

Iaria, A. and A. Wang (2022). The mixed logit and mixed probit are real analytic. *Available at SSRN 4094866*.

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.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Janys, L. and B. Siflinger (2021). Mental health and abortions among young women: Time-varying unobserved heterogeneity, health behaviors, and risky decisions.

Jochmans, K. (2017, 07). Two-Way Models for Gravity. *The Review of Economics and Statistics 99*(3), 478–485.

Jochmans, K. (2018). Semiparametric analysis of network formation. *Journal of Business & Economic Statistics 36*(4), 705–713.

Johnson, E. G. (2004). Identification in discrete choice models with fixed effects. Working paper, Bureau of Labor Statistics.

Ke, Y., J. Li, and W. Zhang (2016). Structure identification in panel data analysis. *The Annals of Statistics 44*(3), 1193–1233.

Kitazawa, Y. (2022). Transformations and moment conditions for dynamic fixed effects logit models. *Journal of Econometrics 229*, 350–362.

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.

Krasnokutskaya, E., K. Song, and X. Tang (2022). Estimating unobserved individual heterogeneity using pairwise comparisons. *Journal of Econometrics 226*(2), 477–497.

Krein, M. G. and A. A. Nudelman (1977). *The Markov Moment Problem and Extremal Problems*. American Mathematical Society.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics 95*(2), 391–413.

Lavergne, P. and V. Patilea (2013). Smooth minimum distance estimation and testing with conditional estimating equations: uniform in bandwidth theory. *Journal of Econometrics 177*(1), 47–59.

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

Lewbel, A. (2014, 02). 38An Overview of the Special Regressor Method. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press.

Lewbel, A. (2019, December). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature 57*(4), 835–903.

Lewbel, A. and S. M. Schennach (2007, January). A simple ordered data estimator for inverse density weighted expectations. *Journal of Econometrics 136*(1), 189–211.

Lewis, D. J., D. Melcangi, L. Pilossoph, and A. Toner-Rodgers (2023). Approximating grouped fixed effects estimation via fuzzy clustering regression. *Journal of Applied Econometrics*.

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, September). Least squares quantization in pcm. *IEEE transactions on information theory 28*(2), 129–137.

Lovász, L. (2012). *Large Networks and Graph Limits.*, Volume 60 of *Colloquium Publications*. American Mathematical Society.

Luo, Z. and P. Tseng (1993, March). Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research 46-47*(1), 157–178. Copyright: Copyright 2007 Elsevier B.V., All rights reserved.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman (Eds.), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, pp. 281–297. University of California Press.

Magnac, T. (2004). Binary variables and sufficiency: Generalizing conditional logit. *Econometrica 72*(6), 1859–1876.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica 5*(2), 357–362.

Manski, C. F. (2021). Econometrics for decision making: Building foundations sketched by haavelmo and wald. *Econometrica 89*(6), 2827–2853.

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

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

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.

Moon, H. R. 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 (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 non-linear panel models with two-way fixed effects. Technical report.

Mundlak, Y. (1961). Empirical production function free of management bias. *American Journal of Agricultural Economics 43*(1), 44–56.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*(1), 1–32.

Ng, A., M. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. In T. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems*, Volume 14. MIT Press.

Pang, J.-S. (1987). A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research 12*(3), 474–484.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica 74*(4), 967–1012.

Pesaran, M. H. (2015, November). *Time Series and Panel Data Econometrics*. Number 9780198759980 in OUP Catalogue. Oxford University Press.

Politis, D. N., J. P. Romano, and M. Wolf (1999). *Subsampling*. Springer Science & Business Media.

Pollard, D. (1981). Strong Consistency of K-Means Clustering. *The Annals of Statistics 9*(1), 135 – 140.

Pollard, D. (1982). A Central Limit Theorem for k-Means Clustering. *The Annals of Probability 10*(4), 919 – 926.

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*. Berlin, – Heidelberg, – New York: Springer.

Roth, J. and Rambachan (2022). A more credible approach to parallel trends. Technical report.

Saggio, R. (2012). Discrete unobserved heterogeneity in discrete choice panel data models. Master's thesis, Center for Monetary and Financial Studies.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 580–615.

Shorack, G. and J. Wellner (1986). *Empirical Processes with Applications to Statistics*. John Wiley and Sons.

Stammann, A. (2018). Fast and feasible estimation of generalized linear models with high-dimensional k-way fixed effects. Technical report.

Stammann, A., F. Heiß, and D. McFadden (2016). Estimating fixed effects logit models with large panel data. Number G01-V3 in Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, Kiel und Hamburg. ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft.

Stukel, T. A. (1988). Generalized logistic models. *Journal of the American Statistical Association 83*(402), 426–431.

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.

Toth, P. (2017). Semiparametric estimation in network formation models with homophily and degree heterogeneity. *Available at SSRN 2988698*.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York, NY.

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.

Whittaker, E. T. and G. N. Watson (1996). *A Course of Modern Analysis* (4 ed.). Cambridge Mathematical Library. Cambridge University Press.

Wooldridge, J. M. (2010). *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. Technical report.

Zhang, Y., E. Levina, and J. Zhu (2017, 09). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika 104*(4), 771–783.

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.

ECOLE DOCTORALE

Titre : Essais en Économétrie des Données de Panel

Mots clés : Données de panel, Modèles non-linéaires, Hétérogénéité inobservée, Identification, Statistiques en grande dimension, Microéconométrie

Résumé : Cette thèse comporte cinq chapitres portant sur l'étude de quelques problèmes d'identification, d'estimation et d'inférence au sein de modèles semi-paramétriques pour l'analyse économétrique des données de panel. Les quatre premiers chapitres se concentrent sur une classe de modèles dits ≪ à effets fixes ≫, où l'hétérogénéité inobservée par l'économètre est approximée par des variables latentes de faible dimension (relativement à la taille des données) dont la distribution conditionnellement aux variables exogènes n'est pas restreinte.

Dans le premier chapitre, nous généralisons un résultat de Johnson (2004) et Chamberlain (2010) en démontrant que l'identification du paramètre de pente, dans un modèle statique de choix discrets avec hétérogénéité individuelle constante dans le temps et des agents observés plus de deux périodes, reste possible hors du cas restrictif où les erreurs suivent une loi logistique. Nous exhibons une restriction sur un moment conditionnel à partir de laquelle un estimateur asymptotiquement normal à vitesse paramétrique, quand le nombre d'individus tend vers l'infini, est obtenu par la méthode généralisée des moments (GMM). Nous illustrons cette nouvelle méthode en revisitant la relation entre déficits budgétaires et réélections étudiée dans Brender et Drazen (2008). L'effet significatif et positif du déficit budgétaire sur la probabilité de réélection est robuste à une relaxation de l'hypothèse logistique.

Dans le second chapitre, nous présentons des conditions d'identification pour une classe de modèles non-linéaires à doubles effets fixes et coefficients hétérogènes séparables lorsque le panel est à la fois long et large. Nous proposons une méthode d'estimation rapide reposant sur une descente de gradient coordonnées par coordonnées exploitant la séparabilité additive des effets fixes. Dans le cas semi-paramétrique, nous démontrons l'équivalence numérique de la méthode avec celle du maximum de vraisemblance, reportons des gains de calcul importants sans perte de précision au regard des méthodes existantes (e.g., logitfe/probitfe dans Stata) et revisitons deux applications empiriques en innovation (Aghion et al., 2013) et commerce international (Helpman et al., 2008). Nous trouvons une hétérogénéité significative des coefficients de pente relatifs aux variables indépendantes dans chacun des modèles.

Les troisièmes et quatrièmes chapitres traitent d'un cas particulier de modèles à facteurs où les coefficients individuels associés à chaque facteur temporel sont supposés discrets. Cette hypothèse génère une structure de groupe qui présente un intérêt dans une variété de situations économiques (e.g., clubs de pays, de partenaires commerciaux, types de consommateurs, produits, actifs financiers). Le troisième chapitre propose un nouvel estimateur en deux étapes pour le modèle linéaire qui présente un certain nombre d'avantages théoriques et computationnels. Grâce à la résolution d'un problème convexe et l'utilisation d'une procédure agglomérative, nous généralisons Bonhomme et Manresa (2015) en montrant que le paramètre de pente, les effets fixes, et le nombre de groupes peuvent être estimés de manière convergente sans borne supérieure connue sur le nombre de groupes, tout en réduisant la complexité algorithmique à l'ordre du nombre d'individus au cube contre une complexité exponentielle pour l'estimateur reposant sur l'algorithme des k-means.

Le quatrième chapitre étend certains de ces résultats à une classe de modèles non-linéaires discrets.

Le cinquième et dernier chapitre démontre la normalité asymptotique d'estimateurs s'exprimant comme la moyenne empirique d'une transformation d'une fonction de répartition empirique par un quantile empirique sous des hypothèses plus faibles que les existantes. Un exemple est l'estimateur ≪ Changes-in-Changes ≫ pour l'effet de traitement moyen proposé dans Athey et Imbens (2006). Des simulations de Monte Carlo suggèrent que nos hypothèses sont difficilement améliorables.

Title : Essays in Panel Data Econometrics

Keywords : Panel data, Nonlinear models, Unobserved heterogeneity, Identification, High-dimensional statistics, Microeconometrics

Abstract : This thesis consists of five chapters dealing with some identification, estimation, and inference problems in a class of semiparametric panel data models for econometric analysis. The first four chapters focus on fixed effects models, in which unobserved heterogeneity (to the econometrician) is approximated by introducing latent variables defined on low-dimensional manifolds (relative to the dimension of the data) and whose distribution conditional on the exogenous variables is left unrestricted.

In the first chapter, we generalize some of Johnson (2004) and Chamberlain (2010)'s results by showing that the slope parameter in a static binary choice model with three periods or more can be point identified even if the idiosyncratic shocks do not follow the restrictive logistic distribution. We provide a conditional moment restriction, which can be used to obtain an asymptotically normal estimator at the parametric rate, when the number of units diverges to infinity, by applying the Generalized Method of Moments (GMM). We illustrate this new method by revisiting the relationship between budget deficits and reelections studied in Brender and Drazen (2008). The significant positive effect of budget deficits on the probability of reelection is robust to departures from the logistic assumption.

In the second chapter, we present identification conditions for a class of nonlinear two-way fixed effects models with heterogeneous coefficients for large and long panels. We provide a fast estimation procedure based on a Gauss-Siedel coordinate-wise gradient descent algorithm which exploits additive separability in the fixed effects. In the semiparametric case, we prove the numerical equivalence of our method to the maximum likelihood estimator, we report considerable gains in execution time without loss in precision with respect to existing packages (e.g., logitfe/probitfe in Stata), and we revisit two empirical applications in innovation (Aghion et al., 2013) and international trade (Helpman et al., 2008). We find significant heterogeneity in estimated slopes for the independent variables in each case.

The third and fourth chapters consider a special case of factor models, in which individual factor loadings are assumed discrete. This assumption generates a group structure that can rationalize a wide variety of economic settings (e.g., clubs of countries, trading partners, types of consumers, goods, financial assets). The third chapter proposes a new two-step estimator for the linear model, which has several theoretical and computational advantages. By solving a convex optimization program and using an agglomerative clustering procedure, we generalize Bonhomme and Manresa (2015) and show that the common slope parameter, the fixed effects, and the number of groups can be consistently estimated without a known upper bound on the number of groups while reducing algorithmic complexity to the order of the cube of the number of units against an exponential complexity for the estimator relying on the k-means algorithm.

The fourth chapter extends some of these results to a class of nonlinear models for discrete outcomes.

The fifth and last chapter proves the asymptotic normality of estimators defined as empirical means of the transform of an empirical cumulative distribution function by an empirical quantile process under much weaker assumptions than what is currently known. One popular example is the "Changes-in-Changes" estimator proposed in Athey and Imbens (2006). Monte Carlo simulations suggest that our assumptions cannot be improved.

Institut Polytechnique de Paris
91120 Palaiseau, France