WARWICK
ECONOMICS

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

Martin Mugnier & Ao Wang

August 2022(Revised September 2022) No: 1422

Warwick Economics Research Papers

ISSN 2059-4283 (online)

ISSN 0083-7350 (print)

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

Martin Mugnier† Ao Wang‡

September 9, 2022

Abstract

We study a nonlinear two-way fixed effects panel model that allows for unobserved individual heterogeneity in slopes (interacting with covariates) and (unknown) flexibly specified link function. The former is particularly relevant when the researcher is interested in the distributional causal effects of covariates, and the latter mitigates potential misspecification errors due to imposing a known link function. We show that the fixed effects parameters and the (nonparametrically specified) 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 (and unobserved) heterogeneity in the causal relationship of interest that may have important implications for the subsequent policy analysis.

*We are grateful to Mingli Chen, Xavier D'Haultfœuille, Christophe Gaillac, Ivana Komunjer, Eric Renault, Amrei Stammann, Francis Vella, and Martin Weidner for their useful suggestions. We also thank the conference participants at the 2021 Bristol Econometric Study Group, 2021 European Winter Meeting of the Econometric Society (ES), Encounters in Econometric Theory at Oxford, 2022 Asia Meeting of the ES in Shenzhen, and seminar participants at CREST, Georgetown, University of Chicago, and Warwick for their helpful comments. Martin Mugnier gratefully acknowledges financial support from the research grants Otelo (ANR-17-CE26-0015-041) and ANR "Investissements d'avenir": EUR DATA EFM (ANR-18-EURE-0005).

†CREST, ENSAE, Institut Polytechnique de Paris, martin.mugnier@ensae.fr

‡University of Warwick, CAGE Research Centre, ao.wang@warwick.ac.uk

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.[1] 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). However, the application of these models is still subject to theoretical and practical challenges. First, the extent to which the model is nonparametrically identified is still 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.[2] 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 paper, we tackle these two challenges in a class of index function static models characterized by the probability of individual $i = 1, ..., N$ from choosing $y_{it} \in \mathcal{Y}$ at time $t = 1, ..., T$:

$$\Pr(y_{it} = y \mid x_{i1}, ..., x_{it}, \alpha_i, \beta_i, \xi_t) = g(y; \alpha_i + \xi_t + x_{it}'\beta_i), \tag{1}$$

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 (unknown) link function. This model encompasses settings with single index, such as binary outcome, ordered outcome and count outcome, as well as those with multiple indices (multimodal outcome), and has been widely used in literature including 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 (1) features two important relaxations. First, we allow the slope parameter, β_i, to be individual-specific rather than common

[1] 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).

[2] For instance, the standard implementation of the MLE requires storing and inverting a Hessian matrix of size 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 3.2 for details.

across individuals ($\beta_i = \beta$).[3] 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 their implications for policy evaluations. Second, the link function g can be left unknown (and therefore to be estimated) rather than specified as a known function (e.g., logit, probit). To the extreme, g can be nonparametric. This flexibility enables to mitigate potential misspecification errors due to parametric choices made on the link function.

The central theoretical question in this paper is then the extent to which the parameters in model (1) are identified under such relaxations. In the setting of large N and large T, we propose a novel identification strategy and prove that β_i, α_i, ξ_t, and function g, can be point-identified. Our strategy crucially relies on the technique of *compensating variable*.[4] 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 then achieving the point identification. Importantly, this strategy does not require strong support condition on the compensating variable such as large support. 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 (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

[3] We can also allow for the slope parameter to be time-specific rather than individual-specific. The main results of the paper still hold. See remark 1 for details.

[4] To the best of our knowledge, this term was first introduced by Hicks (1939) and later appears in Lewbel (2019)'s survey of identification in econometrics. D'Haultfoeuille et al. (2021) use similar arguments to identify common parameters in a single index model with endogeneity and a single instrument (not necessarily) violating the exclusion restriction. See also D'Haultfœuille et al. (2022) (the second statement of Theorem 4.2) for another application of this argument in identification.

individuals (T time periods). This substantially alleviates the computational burden of concentrating out a potentially large number of fixed effects in the routine 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.[5] Second, extensive Monte Carlo simulations suggest its fast convergence. It already achieves a good numerical 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 (1) with multi-way fixed effects.[6] We provide a Python package `nlmfe` that implements this procedure and some of the extensions.[7]

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. However, the residual dispersion in the slopes (i.e., unexplained by the observed characteristics) is still significant. These exercises emphasize the usefulness of the proposed method in capturing flexible (and unobserved) heterogeneity in the causal relationship of interest.

[5] 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 3.2 and Appendix F. Besides, we develop a practical Bootstrap construction of confidence intervals for such features. See Section 4 for the Monte Carlo evidence of its good finite-sample performance.

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

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

Related literature. This paper significantly 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.[8] To the best of our knowledge, this paper is 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. 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 time completely coincides with that of individuals) to achieve identification.[9] We focus on a different setting in which the relationship between time and individual dimensions is unrestricted.[10] 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).[11] In addition, our methodology applies to not only 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 paper 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,

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

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

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

[11] 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 (2017) for examples using this condition in fixed-T setting.

2013; Stammann, 2018; Stammann et al., 2016), EM method (Chen, 2016), or Minorization-Maximation algorithm (Chen et al., 2021b) 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.[12] 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 conditions, ensuring its validity in inference and robust finite-sample performance.

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 $y_{it} \in \mathcal{Y}$:

$$\Pr\left(y_{it} = y \mid x_{i1}, ..., x_{it}, \alpha_i, \beta_i, \xi_t\right) = g(y; \alpha_i + \xi_t + x_{it}'\beta_i), \tag{2}$$

where x_{it} are individual i's observed characteristics at time t, (α_i, β_i) are individual-fixed effects, ξ_t is a (vector of) time-fixed effect(s), and g is a (unknown) link function of outcome y and indices $\alpha_i + \xi_t + x_{it}'\beta_i$. Model (2) covers the case of single index, i.e., $\alpha_i + \xi_t + x_{it}'\beta_i$ is a scalar, as well as multiple indices (and therefore multimodal outcomes), i.e., $\alpha_i + \xi_t + x_{it}'\beta_i$ is a vector. It 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.[13] Second, the link function g in model (2) can be nonparametrically specified, relaxing the

[12] See page 16 in their paper.

[13] 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) with common slopes γ across individuals and time periods.

common parametric restrictions such as probit and logit in empirical research.

Remark 1. *One can consider a model with time-specific slope parameters:*

$$\Pr\left(y_{it} = y \mid x_{i1}, ..., x_{it}, \alpha_i, \beta_t, \xi_t\right) = g(y; \alpha_i + \xi_t + x_{it}'\beta_t). \tag{3}$$

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) *in the main text and extend our main results to model* (3) *in Appendix D.*

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

Example 1 (Binary outcome)**.**

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

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

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

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 $d_2 > d_1$, $(x_{i1}', ..., x_{it}', \alpha_i, \beta_i', \xi_t)'$ 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 $\mathcal{Y} = \{0, 1, 2, ...\}$, model* (2) *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)**.**

$$y_{it} = \underset{j=1,...,J}{\arg\max} \left\{ \alpha_{ij} + \xi_{tj} + x'_{tj}\beta_{ij} - u_{itj} \right\}, \tag{4}$$

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

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

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

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). To simplify the exposition, we present the arguments for $x_{it} = (x_{it}^{(1)}, x_{it}^{(2)}) \in \mathcal{X} \subset \mathbf{R}^2$ and the case of single index in the main text. The results for the case of multimodal outcome (Example 4) and model (3) (heterogeneous slope parameters across time) are presented in Appendices C and D, respectively. Denote by $\mathcal{X}^1$ and $\mathcal{X}^2$ the support of $x_{it}^{(1)}$ and $x_{it}^{(2)}$, respectively.[14] Without loss of generality, suppose that $|\mathcal{X}^1|, |\mathcal{X}^2| > 1$ and normalize $\alpha_1 = 0$, $\xi_1 = 0$, and $\beta_1^{(1)} = 1$.[15]

3.1 Identification

In this section, we assume that both the number of individuals in $\mathbf{N}$ and that of time periods in $\mathbf{T}$ are large. To start with, define a random variable:

$$z_i(x^{(1)}; x^{(2)}) = \alpha_i + \beta_i^{(1)} x^{(1)} + x^{(2)}(\beta_i^{(2)} - \beta_1^{(2)}) \tag{5}$$

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: $\alpha_1 +$

[14] Allowing $\mathcal{X}^1, \mathcal{X}^2$ to depend on i is straightforward but requires heavier notation from which we abstract. In contrast, dependence on t is ruled out as we rule out non-stationary covariates.

[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 $\beta_1^{(1)} = 1$.

$\xi_t + \beta_1^{(1)} z_i(x^{(1)}; x^{(2)}) + \beta_1^{(2)} x^{(2)} = \alpha_i + \xi_t + \beta_i^{(1)} x^{(1)} + \beta_i^{(2)} x^{(2)}$. Define $\mathcal{Z}$ as the closure of $\left\{(z_i(x^{(1)}; x^{(2)}), x^{(2)}) : i \in \mathbf{N}, (x^{(1)}, x^{(2)}) \in \mathcal{X}\right\}$. We propose the following assumptions for identification.

Assumption 1.

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

(ii). *(a) For any given $i \geq 1$, 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 $\tau_t \leq C\rho^t$.*

(b) For any given $t \geq 1$, conditional on ξ_t, $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i \geq 2}$ are independent.

(iii). *For all $(i, i', t) \in \mathbf{N}^2 \times \mathbf{T}$, $\xi_t \perp\!\!\!\perp (\alpha_i, \beta_i, x_{it}^{(1)}) \mid x_{it}^{(2)}$, $\xi_t \mid \left\{x_{it}^{(2)} = x^{(2)}\right\} \stackrel{d}{=} \xi_t \mid \left\{x_{i't}^{(2)} = x^{(2)}\right\} \sim F_\xi(\xi; x^{(2)})$, and $Supp(x_{it} \mid \alpha_i, \beta_i, \xi_t) = \mathcal{X}$.*

(iv). *For all $x^{(2)} \in \mathcal{X}^2$ and $i \in \mathbf{N}$, there exist $x^{(1)}, x^{(1')} \in \mathcal{X}^1$, $x^{(1)} \neq x^{(1')}$, such that $(z_i(x^{(1)}; x^{(2)}), x^{(2)}), (z_i(x^{(1')}; x^{(2)}), x^{(2)}) \in \mathcal{X}$.*

(v). *The level set $\left\{(z, x^{(2)}) \in \mathcal{Z} : z + \beta_1^{(2)} x^{(2)} = r\right\}$ is not a singleton for some $r \in \mathbf{R}$.*

(vi). *For all $t \in \mathbf{T}$, $\left\{z + \beta_1^{(2)} x^{(2)} + \xi_t : (z, x^{(2)}) \in \mathcal{Z}\right\} \cap \left\{z + \beta_1^{(2)} x^{(2)} : (z, x^{(2)}) \in \mathcal{Z}\right\} \neq \emptyset$.*

Remark 2. *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 1(ii)b. Then, the independence among $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i \geq 2}$ (with such covariates being excluded from x_{it}) (as well as our main results below) still holds. See footnote 37 for more details.*

Assumption 1(i) is standard in index function models and satisfied in the examples we provided.[16] Assumption 1(ii) imposes dependence restrictions across individual and time dimensions of the panel. Assumption 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 1(ii)b requires the conditional cross-sectional independence across individuals. Both requirements are standard in the panel data literature.[17] Assumption 1(iii) requires the exogeneity of individual-fixed effect (α_i, β_i) and variable(s) $x_{it}^{(1)}$ with

[16] 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 1′(i) for details.

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

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 1(iv)-(vi) characterize properties of the compensating variable $z_i(x^{(1)}; x^{(2)})$. Assumption 1(iv) specifies the condition under which $z_i(x^{(1)}; x^{(2)})$ compensates between two individuals, i.e., the index of individual i with $(x^{(1)}, x^{(2)})$ and that of individual 1 with $(z_i(x^{(1)}; x^{(2)}), x^{(2)})$ are equal. Intuitively, it will be employed to show the identification of individual-specific fixed parameters α_i, $\beta_i^{(1)}$, and $\beta_i^{(2)} - \beta_1^{(2)}$. Assumption 1(v) 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 prove the identification of $\beta_1^{(2)}$ (and therefore $\beta_i^{(2)}$). It is straightforward to extend this assumption to allow for more than two covariates. Assumption 1(vi) describes the condition under which $z_i(x^{(1)}; x^{(2)})$ compensates between time periods, i.e., the sets of indices (and their limiting points) in time periods t and 1 overlap. It will be used to identify (relative) time-specific fixed effect ξ_t. These three assumptions can be achieved by having a regressor $x^{(1)}$ with large support, e.g., $\mathcal{X}^1 = \mathbf{R}$, but this large support condition is not necessary.[18]

The next theorem summarizes our main identification result:

Theorem 1. *Suppose that Assumptions 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 1(v)-(vi) further hold, then*
  - ξ_t *and* $\beta_i^{(2)}$ *are identified for* $i \in \mathbf{N}$ *and* $t \in \mathbf{T}$.
  - $g(y; v)$ *is identified as a function of* $y \in \mathcal{Y}$ *and index* $v = \alpha_i + \xi_t + x'\beta_i$.

Proof. See Appendix A. □

According to Theorem 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 5, we illustrate this point by revisiting two classic empirical studies. Moreover, Theorem 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

[18] Concretely, it depends on the support of $x^{(2)}$ and the fixed effects. For instance, if the support of $x^{(2)}$, (α_i, β_i), and ξ_t are compact, say, $[-1, 1]$, then it suffices that $x^{(1)}$ varies within $[-3, 3]$ to cover the whole range of $\beta_i^{(2)} x_{it}^{(2)} + \alpha_i + \beta_t$.

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 1 can be extended to a model (2) with multimodal outcomes and model (3) with heterogeneous slopes β_t.[19] See Appendices C and D for details, respectively.

3.2 Estimation

In this section, we propose a convenient iterative estimation procedure of model (2). 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, and sieve estimation of the link function. To simplify the exposition, we focus on a semi-parametric estimation of model (2) with a known g and $\beta_i = \beta$ in the main text. We discuss the extensions to the settings with unknown g and/or heterogeneous slope parameters in Remarks 3 and 4.

Oftentimes, researchers estimate model (2) by treating the unobserved individual and time effects as parameters to be estimated and using a concentrated MLE. Denote the log-likelihood function by

$$\mathscr{L}_{NT}(\theta) := \sum_{i=1}^{N}\sum_{t=1}^{T} \log g(y_{it}; \alpha_i + \xi_t + x_{it}'\beta), \tag{6}$$

where $\theta = (\alpha_2, ..., \alpha_N, \xi_1, ..., \xi_T, \beta)$ 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 $i = 2, ..., N$ and $t = 1, ..., T$:

$$(\widehat{\alpha}_2, ..., \widehat{\alpha}_N, \widehat{\xi}_1, ..., \widehat{\xi}_T) \in \underset{(\alpha_2,...,\alpha_N)\in\mathcal{A},(\xi_1,...,\xi_T)\in\Xi}{\arg\max} \mathscr{L}_{NT}(\alpha_2, ..., \alpha_N, \xi_1, ..., \xi_T, \beta), \tag{7}$$

where $\mathcal{A} := \mathcal{A}_2 \times ... \times \mathcal{A}_N \subset \mathbf{R}^{N-1}$ and $\Xi := \Xi_1 \times ... \times \Xi_T \subseteq \mathbf{R}^T$, 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 (6), one maximizes $\mathscr{L}_{NT}(\theta)$ with

[19] See Dubois et al. (2020) for an application of such multi-index two-way fixed effects models in demand estimation.

respect to β:

$$\widehat{\beta} \in \arg\max_{\beta \in \mathcal{B}} \mathscr{L}_{NT}(\widehat{\alpha}_2, ..., \widehat{\alpha}_N, \widehat{\xi}_1, ..., \widehat{\xi}_T, \beta), \tag{8}$$

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

This standard implementation can be computationally intensive due to two reasons. First, concentration step (7) 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 (8) 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 (7) 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 therefore solving the numerical challenge in concentration step (7). This is doable due to the two-way fixed effects structure in (7): 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 minimization 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 minimization 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 (8).

We provide two algorithms for applied researchers to use in different settings, 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 $(\xi_1^{(0)}, ..., \xi_T^{(0)}, (\beta^{(0)})')' \in \Xi \times \mathcal{B}$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all

$j \in \{1, 2, ...\}$. Set $s = 0$.

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

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

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

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

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 $(\alpha_2^{(0)}, ..., \alpha_N^{(0)}, \xi_1^{(0)}, ..., \xi_T^{(0)}, (\beta^{(0)})')' \in \mathcal{A} \times \Xi \times \mathcal{B}$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, ...\}$. 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 and does not involve any $\alpha_r^{(s)}$ for $r \neq i$ ($\xi_r^{(s)}$ for $r \neq t$).[20] 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 3. *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 Appendix E.1.*

Remark 4. *Both FPMLE and FPMLE^{++} can be easily customized to estimating a unknown link function g. It suffices to additionally update the corresponding parameters in g between Steps 4 and 5 of each algorithm. We provide more details in Appendix E.2.*

Remark 5. *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.*

[20] This is true as long as $\mathcal{A}_i$ and Ξ_t are "nice" convex sets projecting onto is easy, e.g., boxes.

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 $\mathbf{R}^{K+N+T-1}$:

$$\widehat{\theta}_{NT}^{\text{MLE}} = \underset{\theta \in \mathbf{R}^{K+N+T-1}}{\arg\max} \ \mathscr{L}_{NT}(\theta). \tag{9}$$

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.

(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 $\mathbf{R}^{K+N+T-1}$. 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 $\mathbf{R}^{K+N+T-1}$.*

Assumption 2(i) is a standard condition for deriving properties of M-estimators. Assumption 2(ii) features smoothness, concavity and coercivity properties of $\mathscr{L}_{NT}$ that together ensure that problem (9) 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(i)-(ii) are sufficient for the numerical equivalence of FPMLE. We further need Assumption 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(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 guarantee the convergence of FPMLE^{++}.

We now state the numerical equivalence of FPMLE and FPMLE^{++} to the MLE.

Theorem 2.

- *Suppose that Assumptions 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(iii) further holds and* $\nu^{(s)} \equiv \nu$ *is constant such that* $0 < \nu < 1/\bar{L}$ *for some absolute constant* $\bar{L} > 0$*,*[21] *then the sequence of iterates generated by FPMLE*$^{++}$*,* $\left\{\widehat{\theta}_{NT}^{++(s)}\right\}_{s=1,2,\ldots}$*, converges to* $\widehat{\theta}_{NT}^{\text{MLE}}$*.*

Proof. See Appendix B. □

Remark 6. *The numerical convergences of both algorithms still hold in the presence of heterogeneous slopes* β_i*. We refer to the proof in Appendix B for such extension.*

Remark 7. *When the concavity requirement in Assumption 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 unknown and to be estimated. 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 Proposition 2 of Appendix E.2, we show that if FPMLE/FPMLE*$^{++}$ *converges numerically, then it converges to a stationary point of the likelihood function* (6)*. 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 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 Section 4, we use Monte Carlo simulations to investigate how large s is required to guarantee good numerical approximation.

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. 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 method 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

[21] For the definition of $\nu^{(s)}$, see Algorithm FPMLE^{++}. The constant $\bar{L}$ is implicitely defined in the proof of Theorem 2, eq. (B.9). In practice, choosing ν and deriving an upper bound on $\bar{L}$ is straightforward given knowledge of g, Θ_{NT}, and the data.

known link function. To be self-contained, we provide a consistency result for the MLE estimators of the slopes 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.[22] 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 Appendix G. 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]

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 sufficient 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 by using $\widehat{\theta}_{NT}^{\text{MLE}}$. 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):

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

[22] See Appendix F for details.

[23] As detailed in Appendix F, 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 $N^{3/8}$. In theory, it is yet to show the plug-in estimators are $\sqrt{N}$-Gaussian (potentially subject to an asymptotic bias) to validate the proposed Bootstrap inference procedure. While our Monte Carlo simulations in the next section seem to support this theoretical statement (at least for the mean of β_i), we leave it 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 $\sqrt{N}$).

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 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 1(iii)). Besides, DGP (iv) violates the stationary requirement in Assumption 1(ii)a.

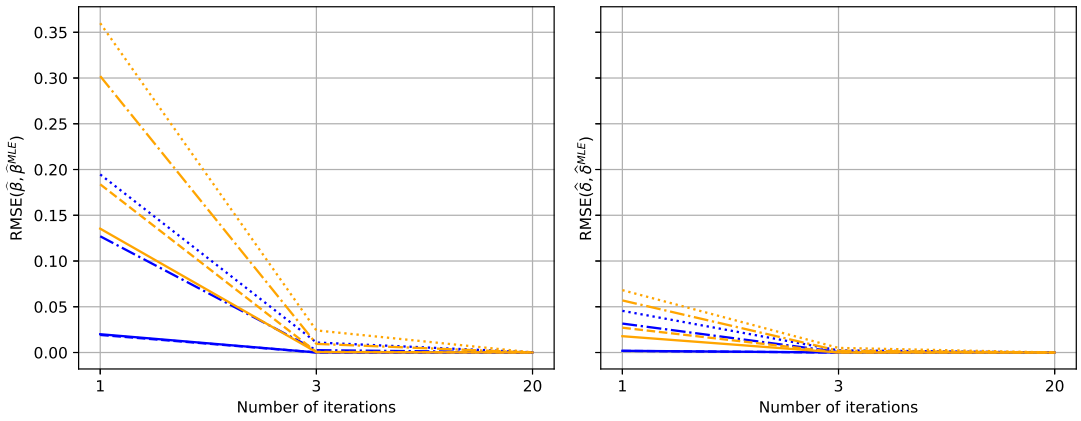
Calibration of the number of iterations. Figure 1 summarizes the "Root Mean Squared Error" (RMSE) distance to the MLE estimator for $\widehat{\beta}_{NT}$ (blue, left panel), $\widehat{\beta}^{++}_{NT}$ (orange, left panel), and the corresponding estimated APEs $\widehat{\delta}_{NT}$ and $\widehat{\delta}^{++}_{NT}$ (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 G.2.

As the number of iterations increases, both FPMLE and FPMLE^{++} converge to the MLE as predicted by Theorem 2. FPMLE already 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. Moreover, the RMSE distance for $\widehat{\delta}_{NT}$ is even a magnitude smaller. For a given number of iterations, FPMLE^{++} approximates less well than FPMLE. Surprisingly, for the DGPs we consider, FPMLE^{++} with 20 iterations already achieves comparable precision to FPMLE (see Table G.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 G.1.

Reduced execution time by FPMLE and FPMLE^{++}. Figure 2 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 the main text and calibrate the number

[24] For the definitions of the RMSE distance to the MLE estimator $\widehat{\theta}^{\text{MLE}}$ and $\widehat{\delta}_{NT}$, see Eq. (G.1) and (G.2), respectively.

Figure 1: 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.*

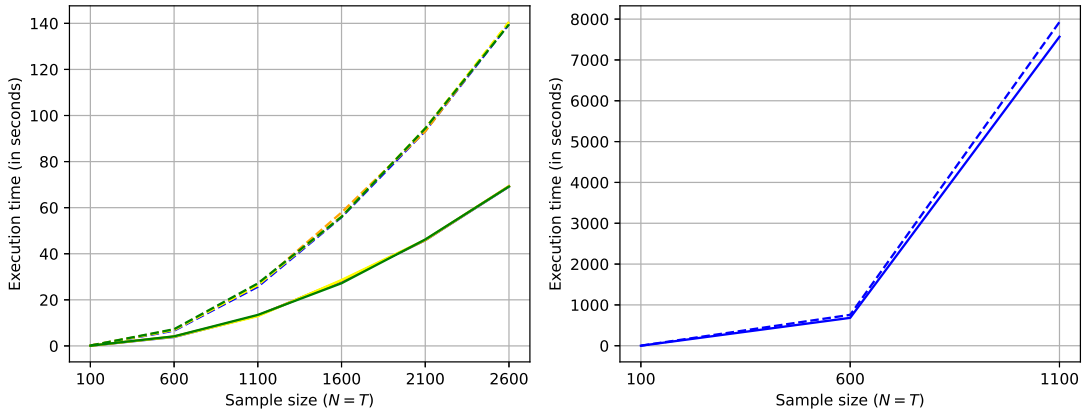
of iterations to be 20 (and jointly with the usual numerical stopping criteria).[25] Overall, FPMLE^{++} largely outperforms `logitfe` in terms of execution time for all sizes of data sets. First, whether or not implementing post-estimation bias correction (dotted lines in Figure 2), FPMLE^{++} only 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 around 1min. This remarkable time efficiency remains valid when $N \gg T$. See Table G.1 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

[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: Execution Time (in Seconds), 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).*

as the APEs.[27] The details of the Poisson model and the implementation of the Bootstrap procedure are included in Appendix G.2.

Table 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 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 $\sqrt{\text{Var}(\beta_{0i})}$ 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 G.4 for details.

[27] In practice, when T or N is moderate, 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 which could serve as a practical alternative. As expected, for DGPs (i) and (ii) that satisfy Assumption 1, the corresponding coverages are dominated by the procedure with the jacknife correction. In contrast, for DGPs (iii) and (iv) that violate Assumption 1, we do not find such dominance. Another potential method is analytical bias correction, which we leave for future research.

Table 1: Inference – Poisson Model with Heterogeneous Slopes

	$\widehat{\mathbb{E}}(\beta_{0i})$			$\sqrt{\widehat{\mathrm{Var}}(\beta_{0i})}$			$\widehat{\mathrm{APE}}$		
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 G.2 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.

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 commonly used nonlinear models (e.g., logit, probit, Poisson), Assumption 2 is satisfied and the likelihood function has a unique global maximum. Then, both algorithms converge to the desired solution. When Assumption 2 may not hold (e.g., when the link function is unknown), the likelihood function may have multiple local maxima. One can then use both algorithms to fast back out the set of stationary points of the likelihood function (Proposition 2 in Appendix E.2) and find out the global maximum, solving the practical challenge of multiple local maxima in the MLE approach (see also remark 7). 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). 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^{++}.

5 Empirical Illustrations

In this section, 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 the complexity in the underlying mechanism(s).

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):

$$\Pr(T_{ij} = 1 \mid \text{dist}_{ij}, \phi_{ij}, \zeta_i, \xi_j) = \Phi\left(-\gamma \text{dist}_{ij} + \phi_{ij}\kappa + \zeta_i + \xi_j\right), i, j = 1, ..., N, i \neq j, \quad (10)$$

where $T_{ij} = 1$ when country j exports to i and zero otherwise, dist_{ij} is the distance between i and j, ϕ_{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:

$$\Pr(T_{ij} = 1 \mid \text{dist}_{ij}, \phi_{ij}, \zeta_i, \xi_j) = \Phi(-\gamma_j^{\text{exp}} \text{dist}_{ij} + \phi_{ij}\kappa + \zeta_i + \xi_j), i, j = 1, ..., N, i \neq j. \quad (11)$$

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 (11) 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:

$$\Pr(T_{ij} = 1 \mid \text{dist}_{ij}, \phi_{ij}, \zeta_i, \xi_j) = \Phi(-\gamma_i^{\text{imp}} \text{dist}_{ij} + \phi_{ij}\kappa + \zeta_i + \xi_j), i, j = 1, ..., N, i \neq j. \quad (12)$$

Similar to (11), the specification in (12) 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, (12) 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. (2008) using (11) and (12), and quantify the extent to which the trade elasticity is heterogeneous among countries.[29]

We estimate (11) and (12) 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 γ_j^{exp} and 158 estimated γ_i^{imp}. The average of $-\gamma_j^{\text{exp}}$ (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 $-\gamma_j^{\text{exp}}$ and $-\gamma_i^{\text{imp}}$ (Figures 3(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

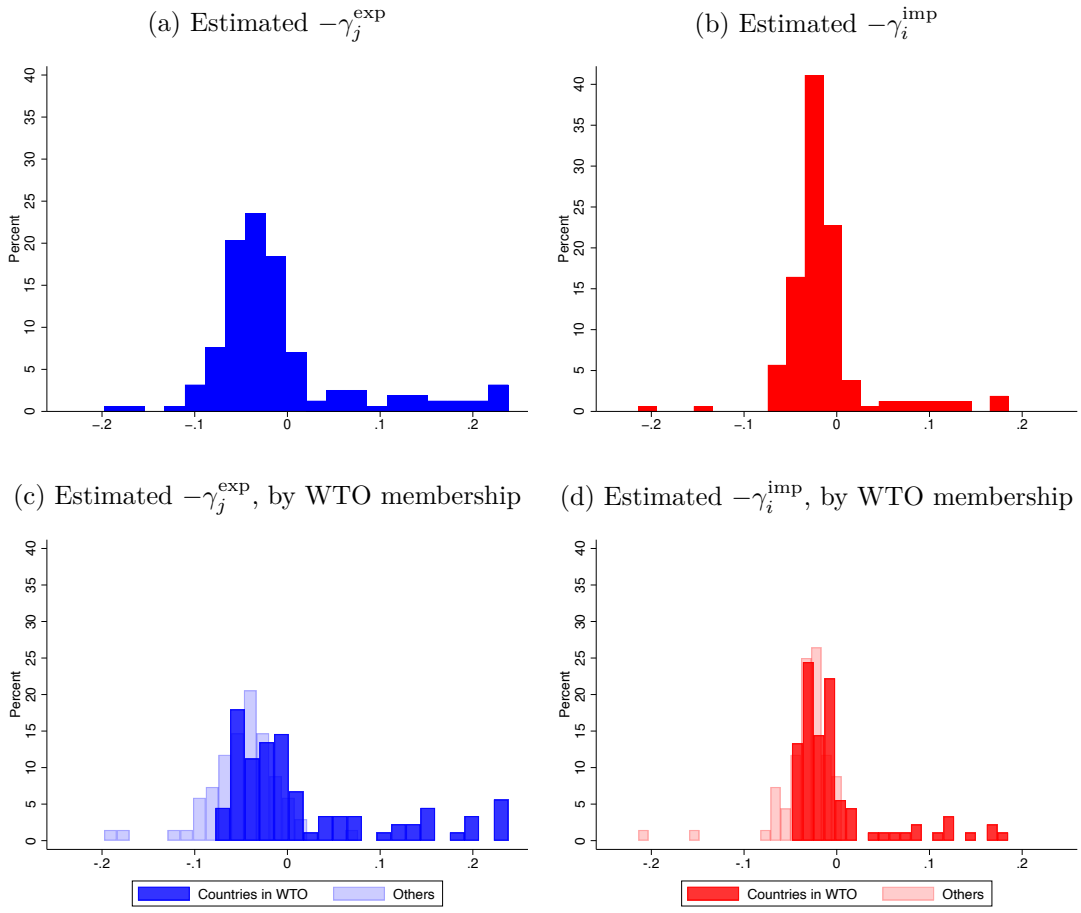
[28] Concretely, denote by ε_j the demand elasticity in country j in their equation 2 on page 449. Then, the log of trade cost, $\ln \tau_{ij}$, enters the first (and the second) step with a coefficient $\varepsilon_j - 1$. As a result, along the lines of their empirical specification, we can specify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

[29] Allowing for country-specific trade elasticity such as (11) and (12) 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.

statistically significant.[31] This dispersion can be partly explained by country characteristics that may determine the trade elasticity. In Figures 3(c) and (d), we plot the distribution of $-\gamma_j^{\text{exp}}$ 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 3(c) and (d). See also Table H.5). The residual dispersion after controlling for such observed characteristics does not seem to disappear. We implement linear regressions of $-\gamma_j^{\text{exp}}$ 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 H.5.

Figure 3: Distribution of Estimated Trade Elasticity



Notes: These histograms are based on 157 *and* 158 *estimated* $-\gamma_j^{exp}$ *and* $-\gamma_i^{imp}$*, respectively.*

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

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:

$$\begin{aligned} &\text{CITES}_{it} \sim \text{Poisson}(\lambda_{it}), \\ &\lambda_{it} = \exp\left\{\beta \times \text{INSTIT}_{it} + \alpha \mathbf{x}_{it} + \eta_i + \tau_t\right\}, \end{aligned} \tag{13}$$

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 (13).[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).

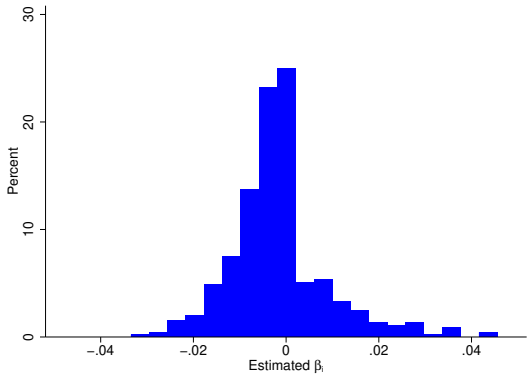
Figure 4 summarizes the unconditional distribution of $\widehat{\beta}_i$ (panel a) and some conditional distributions (panels b and c). The estimated average of β_i is very close to zero (-0.0017). In addition, firms with greater R&D investment (panel b) and Tobin'Q (pabel c) are estimated to react 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).[33] We then decompose $\widehat{\beta}_i$ over a set of firm i's characteristics, denoted

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

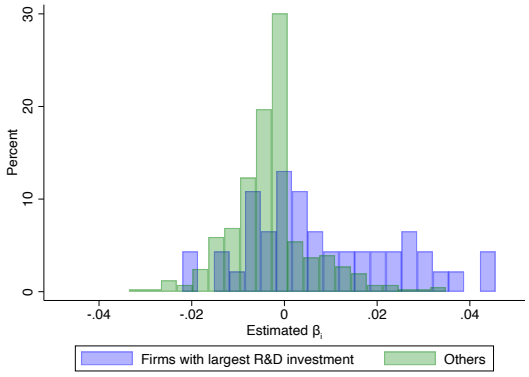
[33] In their Table 1, estimated β ranges from 0.005 to 0.01.

Figure 4: Distributions of $\widehat{\beta}_i$

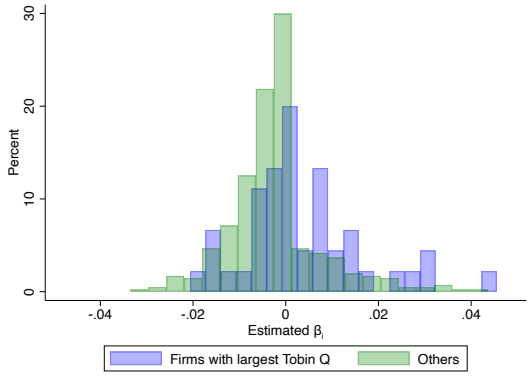
(a) Unconditional



(b) Conditional on R&D investment



(c) Conditional on Tobin's Q



Notes: These histograms are 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.*

by z_i, as follows:

$$\widehat{\beta}_i = z_i\gamma^{\beta} + \zeta_i^{\beta}. \tag{14}$$

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}.[34] we find that ζ_i^{β} still explains around 62% of the total variation in $\widehat{\beta}_i$. See Table H.6 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 correla-

[34]We include the average of sales, R&D expenditure, Tobin's Q across time period, and sector dummies in z_i.

tion, we decompose $\widehat{\eta}_i$ over the same set of firm i's characteristics z_i in (14):

$$\widehat{\eta}_i = z_i\gamma^\eta + \zeta_i^\eta. \tag{15}$$

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%.[35]

6 Conclusion

We study a class of nonlinear two-way fixed effects panel models that features individual-specific slopes (interacting with covariates) in addition to the usual individual-specific and time-specific intercepts, and a unknown (and flexibly specified) link function. The former is particularly 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 assuming a known 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 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.

[35] See Tables H.6 and H.7 for more details of the decomposition.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67(2), 251–333.

Aghion, P., J. Van Reenen, and L. Zingales (2013): "Innovation and institutional ownership," *American economic review*, 103(1), 277–304.

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.

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

Beck, A., and L. Tetruashvili (2013): "On the Convergence of Block Coordinate Descent Type Methods," *SIAM Journal on Optimization*, 23(4), 2037–2060.

Bergé, L. (2018): "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm," Discussion paper, Department of Economics at the University of Luxembourg.

Berry, S., A. Gandhi, and P. Haile (2013): "Connected substitutes and invertibility of demand," *Econometrica*, 81(5), 2087–2111.

Bertsekas, D. (2016): *Nonlinear Programming*, Athena scientific optimization and computation series. Athena Scientific, 3^{rd} edn.

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.

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

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. (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," Discussion paper.

Chen, M., I. Fernández-Val, and M. Weidner (2021a): "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 (2021b): "Payment Instrument Choice with Scanner Data: An MMAlgorithm for Fixed Effects in Multinomial Logit Models," Discussion paper.

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.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013): "Average and Quantile Effects in Nonseparable Panel Models," *Econometrica*, 81(2), 535–580.

Correia, S., P. Guimarães, and T. Zylkin (2020): "Fast Poisson estimation with high-dimensional fixed effects," *The Stata Journal*, 20(1), 95–115.

Czarnowske, D., and A. Stammann (2020): "Fixed Effects Binary Choice Models: Estimation and Inference with Long Panels," .

de Paula, A. (2020): "Econometric Models of Network Formation," *Annual Review of Economics*, 12(1), 775–799.

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., 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*.

Dubois, P., R. Griffith, and M. O'Connell (2020): "How well targeted are soda taxes?," *American Economic Review*, 110(11), 3661–3704.

Evdokimov, K. (2010): "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity," .

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

Fernández-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.

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

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*, pp. 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, W. Y. (2020): "Nonparametric identification in index models of link formation," *Journal of Econometrics*, 215(2), 399–413.

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'.

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

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.

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 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.

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.

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.

Hospido, L. (2012): "Estimating Nonlinear Models with Multiple Fixed Effects: A Computational Note*," *Oxford Bulletin of Economics and Statistics*, 74(5), 760–775.

Iaria, A., and A. Wang (2021): "An Empirical Model of Quantity Discounts with Large Choice Sets," *Working Paper*.

Jochmans, K. (2018): "Semiparametric analysis of network formation," *Journal of Business & Economic Statistics*, 36(4), 705–713.

Lancaster, T. (2000): "The incidental parameter problem since 1948," *Journal of econometrics*, 95(2), 391–413.

Lewbel, A. (2019): "The Identification Zoo: Meanings of Identification in Econometrics," *Journal of Economic Literature*, 57(4), 835–903.

Luo, Z., and P. Tseng (1993): "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.

Neyman, J., and E. L. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica: Journal of the Econometric Society*, pp. 1–32.

Pang, J.-S. (1987): "A Posteriori Error Bounds for the Linearly-Constrained Variational Inequality Problem," *Mathematics of Operations Research*, 12(3), 474–484.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer Science & Business Media.

Shen, X., and W. H. Wong (1994): "Convergence rate of sieve estimates," *The Annals of Statistics*, pp. 580–615.

Stammann, A. (2018): "Fast and Feasible Estimation of Generalized Linear Models with High-Dimensional k-way Fixed Effects," Discussion paper.

Stammann, A., F. Heiss, and D. McFadden (2016): "Estimating Fixed Effects Logit Models with Large Panel Data," No. 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.

Toth, P. (2017): "Semiparametric estimation in network formation models with homophily and degree heterogeneity," *Available at SSRN 2988698*.

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

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

Appendix

Notation: For any $p \geq 1$, any two vectors x and y in $\mathbf{R}^p$, 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 $f : \mathbf{R}^p \to \mathbf{R}$, we let $\nabla f(x)$ (resp. $\nabla^2 f(x)$) denotes its gradient (resp. Hessian) at $x \in \mathbf{R}^p$. For a matrix A, we denote A' as the transpose of A. For a real symetric matrix $A \in \mathbf{R}^{n \times n}$, we let $\lambda_1(A) \geq ... \geq \lambda_n(A)$ denote its real eigenvalues. For any real matrix $A \in \mathbf{R}^{n \times m}$, $\|A\|_2 := \sqrt{\lambda_1(A'A)}$ 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 $\|A\|_{\max} := \max_{i=1,...,n;j=1,...,m} |A_{ij}|$ denotes the max norm.

A Proof of Theorem 1

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

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

Then, under Assumption 1(ii), we can identify $\Gamma_i(\overline{y}; x)$ for each $x \in \mathcal{X}$. When $\mathcal{X}$ is discrete, the identification is achieved by using the law of large numbers (LLN) and Slutsky's lemma.[36] When x_{it} has continuous components, it can achieved by using Nadaraya and Watson's estimator under standard regularity conditions on the density of x_{it} (see, e.g. Hansen, 2008). Using Assumption 1(iii), we obtain:[37]

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

Similarly,

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

[36] 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 $\mathbf{1}\left\{y_{it} = \overline{y}\right\}$ across countably many time periods as long as the correlation between $\mathbf{1}\left\{y_{it} = \overline{y}\right\}\mathbf{1}\left\{x_{it} = x\right\}$ and $\mathbf{1}\left\{y_{it'} = \overline{y}\right\}\mathbf{1}\left\{x_{it'} = x\right\}$ decreases to zero when $|t - t'| \to \infty$.

[37] 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 $x \in \mathcal{X}$. Fixing $x^{(2)}$, Assumption 1(iv) ensures that we can find (and identify) $x^{(1)}$ and $\widetilde{x}^{(1)}$ such that $(x^{(1)}, x^{(2)}), (\widetilde{x}^{(1)}, x^{(2)}) \in \mathcal{X}$ and

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

In particular, it holds for $\widetilde{x}^{(1)} = z_i(x^{(1)}; x^{(2)})$. Because of the monotonicity in Assumption 1(i) and the normalization $\beta_1^{(1)} = 1$, $\Gamma_1(\overline{y}; (z, x^{(2)}))$ is strictly monotonic with respect to z. Then, by inverting $\Gamma_1(\overline{y}; (\cdot, x^{(2)}))$,

$$\widetilde{x}^{(1)} = \Gamma_1^{-1}(\overline{y}; (\Gamma_i(\overline{y}; (x^{(1)}, x^{(2)})), x^{(2)}))$$

is uniquely defined. Hence, we identify $z_i(x^{(1)}; x^{(2)})$. We then apply the same argument to $x^{(1')}$ in Assumption 1(iv) and identify $z_i(x^{(1')}; x^{(2)})$. Using the definition of $z_i(x^{(1)}; x^{(2)})$ and $z_i(x^{(1')}; x^{(2)})$, we identify $\beta_i^{(1)}$ and $\alpha_i + (\beta_i^{(2)} - \beta_1^{(2)})x^{(2)}$. Applying the same reasoning to $x^{(2')} \neq x^{(2)}$, we then separately identify α_i and $\beta_i^{(2)} - \beta_1^{(2)}$.

Second, we identify $\beta_i^{(2)}$ and ξ_t by further using Assumptions 1(ii), (v), and (vi). Note that $z_i(x_{it}^{(1)}; x_{it}^{(2)})$ is already identified from the arguments in the previous paragraph for any $(x_{it}^{(1)}; x_{it}^{(2)}) \in \mathcal{X}$. Given t and conditional on (ξ_t, β_1), because of the independence of $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i \geq 1}$ in Assumption 1(ii), $\{(y_{it}, z_i(x_{it}), x_{it})\}_{i \geq 1}$ are also independent. Then, we identify the following quantity (similarly to that of $\Gamma_i(\overline{y}; x)$):

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

Using the definition of $z_i(x_{it})$, we have

$$\Gamma^t(y; z, x^{(2)}) = g(y; z + \beta_1^{(2)} x^{(2)} + \xi_t)$$

and therefore $g(y; z + \beta_1^{(2)} x^{(2)} + \xi_t)$ is identified for any $y \in \mathcal{Y}$, $(z, x^{(2)}) \in \mathcal{Z}$, and $t \geq 1$. Because of Assumption 1(v), we can find two different pairs $(z, x^{(2)})$ and $(z', x^{(2')})$, such that $x^{(2)} \neq x^{(2')}$ and $g(y; z + \beta_1^{(2)} x^{(2)} + \xi_t) = g(y; z' + \beta_1^{(2)} x^{(2')} + \xi_t)$. Using Assumption 1(i) and setting $y = \overline{y}$, we obtain $z + \beta_1^{(2)} x^{(2)} + \xi_t = z' + \beta_1^{(2)} x^{(2')} + \xi_t$, identifying $\beta_1^{(2)}$ (and therefore $\beta_i^{(2)}$).

According to Assumption 1(vi), we can find two series $(z, x^{(2)})$ and $(z', x^{(2')})$ in $\mathcal{Z}$ such that $z + \beta_1^{(2)} x^{(2)} + \xi_t$ and $z' + \beta_1^{(2)} x^{(2')}$ converge to the same point, denoted by $v^* = z^* + \beta_1^{(2)} x^{(2*)} + \xi_t = z'^* + \beta_1^{(2)} x^{(2'*)}$. Because both $g(\overline{y}; z + \beta_1^{(2)} x^{(2)} + \xi_t)$ and $g(\overline{y}; z + \beta_1^{(2)} x^{(2)})$ are identified, then at v^* and by Assumption 1(i), we obtain $z^* + \beta_1^{(2)} x^{(2*)} + \xi_t = z'^* + \beta_1^{(2)} x^{(2'*)}$ and identify ξ_t.

Finally, because $\{\xi_t\}_{t\geq 1}$ and $\{(\alpha_i, \beta_i)\}_{i\geq 1}$ are identified, we then identify the index $v_{it} = \alpha_i + \xi_t + x'\beta_i$ for any i and t. Then, under Assumption 1(ii), we identify $\mathbb{E}[\mathbf{1}\{y_{it} = y\} \mid v_{it} = v] = g(y; v)$ using the LLN on the support of v_{it}.

B Proof of Theorem 2

B.1 Preliminary results

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

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y \in X. \tag{B.1}$$

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

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in X. \tag{B.2}$$

Lemma 1.

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

$$\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq \mu \|x - y\|^2, \quad \forall x, y \in X. \tag{B.3}$$

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

   (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.*

Proof. 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 (B.3) 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. □

Proposition 1 (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 $X = X_1 \times ... \times X_m$ of closed convex subsets $X_i \subset \mathbf{R}^{n_i}$ such that $\sum_{i=1}^m n_i = n$. Suppose that for each $x = (x_1, ..., x_m) \in X$ and $i \in \{1, ..., m\}$, $y \mapsto f(x_1, ...x_{i-1}, y, x_{i+1}, ..., x_m)$ 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 $x^{k+1} = (x_1^{k+1}, ..., x_m^{k+1})$, given the current iterate $x^k = (x_1^k, ..., x_m^k)$, according to the iteration

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

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

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in X. \tag{B.4}$$

Proof. See p.324 in Bertsekas (2016). □

B.2 Proof of Theorem 2: FPMLE

We proceed in two steps. In a first step, we apply Proposition 1 to $f = -\mathscr{L}_{NT}$ to show that any limit point of the sequence of iterates generated by FPMLE, $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,...}$, 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 $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,...}$ is a stationary point of $-\mathscr{L}_{NT}$. We show that the conditions of Proposition 1 hold for $f = -\mathscr{L}_{NT}$, $m = N + T$, $n_1 = ... = n_{N+T-1} = 1, n_{N+T} = K, n = K + N + T - 1$, $X_1 \times ... \times X_{N-1} = \mathcal{A}_2 \times ... \times \mathcal{A}_N$, $X_N \times ... \times X_{N+T-1} = \Xi_1 \times ... \times \Xi_T$, and $X_{N+T} = \mathcal{B}$. By Assumption 2(i), $X = X_1 \times ... \times X_{N+T} = \Theta_{NT}$ is a Cartesian product of closed convex sets. By Assumption 2(ii), f is continuously differentiable over X. Let $(\alpha, \xi, \beta) \in \Theta_{NT}$, and define the $m = N + T$ real-valued functions

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

We prove that each of the sets

$$\arg\min_{a\in\mathcal{A}_{i+1}} f_{i,\beta,\alpha_{-(i+1)},\xi}(a),\ \arg\min_{e\in\Xi_t} f_{N+t-1,\beta,\alpha,\xi_{-t}}(e),\ \text{and}\ \ \arg\min_{b\in\mathcal{B}} f_{N+T,\alpha,\xi}(b) \quad \text{(B.5)}$$

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 $f_{N+T,\alpha,\xi}$. 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

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

By Assumption 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, $f_{N+T}(b_n) \le \max(m+1,0)$). 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 $\mathbf{R}^K$. As a closed and bounded set of $\mathbf{R}^K$, $\mathcal{B}_0$ is compact in $\mathbf{R}^K$. 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 $b\in\mathcal{B}_0$, $f_{N+T}(b^*)\le f(b)$ by definition of b^*. If $b\notin\mathcal{B}_0$ then, either $m\in\mathbf{R}$ and thus $f_{N+T}(b)\ge m+1\ge f_{N+T}(b^*)$, either $m=-\infty$ and thus $f_{N+T}(b^*)\le 0\le f_{N+T}(b)$. In both cases, $f_{N+T}(b)\ge f_{N+T}(b^*)$.

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(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

$$f_{N+T}(tb_1+(1-t)b_2) = \overline{f}_{N+T}. \quad \text{(B.6)}$$

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 (B.6). 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 $\widehat{\theta}^{\mathrm{MLE}}_{NT}$.

Finally, the monotonicity condition required to apply Proposition 1 easily follows from the strict convexity of f.

Step 2: $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,...}$ 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 $\left\{\widehat{\theta}_{NT}^{(s)}\right\}_{s=1,2,...}$. 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 $\Theta_{NT}^{*} = \left\{\widehat{\theta}_{NT}^{\text{MLE}}\right\}$.

B.3 Proof of Theorem 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 $\theta_1 := \alpha_2, ..., \theta_{N-1} = \alpha_N, \theta_N = \xi_1, ..., \theta_{N+T-1} = \xi_T$, and $\theta_{N+T} = \beta$. Let X_i be the reference space of θ_i (e.g., $X_{N+T} = \mathcal{B}$) and $X = X_1 \times ... \times X_{N+T} = \Theta_{NT}$. 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 $i \in \{1, ..., N+T\}$. 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.[38]

First, Luo and Tseng (1993)'s Assumption A holds because f is convex. Second, by Assumptions 2(i) and 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 1.2, it follows that f is (μ, Θ_{NT})-strongly convex and (L, Θ_{NT})-smooth for some $\mu, L > 0$.[39] Theorem 3.1 in Pang (1987) (whose Assumption (B) holds by Lemma 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 $i \in \{1, ..., N+T\}$,

[38] Note that we do not apply Luo and Tseng (1993)'s Proposition 3.4 for FPMLE because they require more stringent conditions than Proposition 1. 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 $(\mu, \mathbf{R}^{N+T+K-1})$-strong convexity and $(L, \mathbf{R}^{N+T+K-1})$-smoothness of $\mathscr{L}_{NT}$ which rarely holds in our econometric examples, except for rare exceptions (e.g., $(L, \mathbf{R}^{N+T+K-1})$-smoothness of $\mathscr{L}_{NT}$ holds for the logit model).

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

and any $\theta_i' \in X_i$,

$$f(\theta_1, ..., \theta_{i-1}, \theta_i', \theta_{i+1}, ..., \theta_{N+T}) - f(\theta) + \langle \nabla_i f(\theta), \theta_i - \theta_i' \rangle \geq \mu_i \left\| \theta_i' - \theta_i \right\|^2, \quad \text{(B.7)}$$

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

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

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

$$\theta^{(s+1)} = \left[\theta^{(s)} - \nu^{(s)} \nabla f(\theta^{(s)}) + e^{(s)} \right]_X^+, \quad \text{(B.8)}$$

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

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

Under Assumptions 2(i) and 2(iii), there exist positive constants $L_{11}, ..., L_{N+TN+T}$ such that, for all i, j,

$$\left\| \nabla_i f(x) - \nabla_i f(x_1, ..., x_{j-1}, y_j, x_{j+1}, ..., x_{N+T}) \right\| \leq L_{ij} \left\| y_j \right\|, \quad \forall (x, y_j) \in X \times X_j. \quad \text{(B.9)}$$

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 (B.9), we have

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

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

$$\begin{aligned}\left\|e^{(s)}\right\| &\leq \sum_{i=1}^{N+T} \left\|e_i^{(s)}\right\| \\ &\leq \bar{\nu}\bar{L}\left\|\theta^{(s)} - \theta^{(s+1)}\right\| \sum_{i=1}^{N+T} i \\ &\leq \frac{(N+T)(N+T+1)}{2}\left\|\theta^{(s)} - \theta^{(s+1)}\right\|. \end{aligned} \tag{B.10}$$

Equations (B.8) and (B.10) 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 $i \in \{1, ..., N+T\}$,

$$\left\langle \nabla_i f(\theta_1^{(s+1)}, ..., \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, ..., \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \leq 0. \tag{B.11}$$

Equation (3.3.) with $\kappa_2 = \bar{\mu}$ in Luo and Tseng (1993) then immediately follow by summing (B.7) over $i \in \{1, ..., N+T\}$ and cancelling terms:

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

Let us show (B.11). For each $i \in \{1, ..., N+T\}$, Taylor-Lagrange formula with integral remainder ensures

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

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

$$\begin{aligned}&\left\langle \nabla_i f(\theta_1^{(s+1)}, ..., \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, ..., \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\ &\quad = \left\langle \nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\ &\qquad + \left\langle \int_0^1 \nabla^2 f_i(\theta_i^{(s)} + t(\theta_i^{(s+1)} - \theta_i^{(s)}))(\theta_i^{(s+1)} - \theta_i^{(s)})dt, \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle.\end{aligned} \tag{B.12}$$

Eq. (B.9) with $j = i$ implies that f_i is (L_{ii}, X_i)-smooth so that by Lemma 1.2(b), we have

$$\nabla^2 f_i(x_i) \lesssim \bar{L}I, \quad \forall x_i \in X_i. \tag{B.13}$$

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, (B.13), and monotonicity of the integral together, we obtain

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

Plugging this result into (B.12) yields

$$\begin{aligned}
&\left\langle \nabla f_i(\theta_1^{(s+1)}, ..., \theta_i^{(s+1)}, \theta_{i+1}^{(s)}, ..., \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle \\
&\qquad \leq \left\langle \nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}), \theta_i^{(s+1)} - \theta_i^{(s)} \right\rangle + \bar{L} \left\| \theta_i^{(s+1)} - \theta_i^{(s)} \right\|^2 .
\end{aligned} \tag{B.14}$$

Notice that

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

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

$$\begin{aligned}
a(\theta_i^{(s+1)} - \theta_i^{(s)}) &= -a\nu^{(s)} \left[\nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}) \right]_{X_i}^{+} \\
&= -\nu^{(s)} \left[a\nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}) \right]_{X_i}^{+} \\
&= -a\nu^{(s)} \nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}).
\end{aligned} \tag{B.15}$$

Hence, multiplying the first term of (B.14) by a gives

$$\begin{aligned}
&\left\langle \nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}), a(\theta_i^{(s+1)} - \theta_i^{(s)}) \right\rangle \\
&= -a\nu^{(s)} \left\| \nabla_i f(\theta_1^{(s+1)}, ..., \theta_{i-1}^{(s+1)}, \theta_i^{(s)}, ..., \theta_{N+T}^{(s)}) \right\|^2 .
\end{aligned} \tag{B.16}$$

Next, since orthogonal projection is a contraction, using (B.9) and multiplying the

second term of (B.14) by a we obtain

$$a\left\|\theta_i^{(s+1)}-\theta_i^{(s)}\right\|^2 = a\bar{L}\nu^{(s)2}\left\|\left[\nabla_i f(\theta_1^{(s+1)},...,\theta_{i-1}^{(s+1)},\theta_i^{(s)},...,\theta_{N+T}^{(s)})\right]_{X_i}^{+}\right\|^2$$
$$\leq a\bar{L}\nu^{(s)2}\left\|\nabla_i f(\theta_1^{(s+1)},...,\theta_{i-1}^{(s+1)},\theta_i^{(s)},...,\theta_{N+T}^{(s)})\right\|^2. \qquad \text{(B.17)}$$

Finally, multiplying (B.14) by a, combining (B.16)-(B.17), and dividing by a yields

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

Because $0 \leq \nu^{(s)} = \nu \leq 1/\bar{L}$ the right hand-side is negative, which proves (B.11).

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 is complete.

Numerical convergence in the presence of heterogeneous slopes. We need to first strengthen Assumption 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(ii) and the smoothness in Assumption 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 $1/\bar{L}$.

C Extension of Theorem 1 to Multimodal Outcomes

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

$$\Pr(y_{it} = j \mid (\alpha_{ij}, \xi_{tj}, \beta_{ij}, x_{itj})_{j=1}^{J}) = g_j(\delta_{it}), \tag{C.1}$$

where $\delta_{it} = (\delta_{itj})_{j=1}^{J}$ with $\delta_{itj} = \alpha_{ij} + \xi_{tj} + x'_{itj}\beta_{ij}$ and $\sum_{j=1}^{J} g_j(\delta_{it}) < 1$. The residual probability, $g_0(\delta_{it}) = 1 - \sum_{j=1}^{J} g_j(\delta_{it})$, is usually defined as the probability of choosing the outside option. Model (C.1) is a common setting in empirical research such as demand estimation (Berry et al., 2013; Dubois et al., 2020). Define $g(\delta_{it}) = (g_j(\delta_{it}))_{j=1}^{J}$, $\alpha_i = (\alpha_{ij})_{j=1}^{J}$, $\beta_i = (\beta_{ij})_{j=1}^{J}$, $\xi_t = (\xi_{tj})_{j=1}^{J}$. For $j = 1, ..., J$, 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:

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

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

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

We first extend Assumption 1 to the following:

Assumption 1′.

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

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

(ii). (a) For any given $i \geq 1$, 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 $\tau_t \leq C\rho^t$.

(b) *For any given* $t \geq 1$, *conditional on* ξ_t, $\{(y_{it}, x_{it}, \alpha_i, \beta_i)\}_{i \geq 2}$ *are independent.*

(iii). *For all* $(i, i', t) \in \mathbf{N}^2 \times \mathbf{T}$, $\xi_t \perp\!\!\!\perp (\alpha_i, \beta_i, x_{it}^{(1)}) \mid x_{it}^{(2)}$, $\xi_t \mid \left\{x_{it}^{(2)} = x^{(2)}\right\} \stackrel{d}{=} \xi_t \mid \left\{x_{i't}^{(2)} = x^{(2)}\right\} \sim F_\xi(\xi; x^{(2)})$, *and* $Supp(x_{it} \mid \alpha_i, \beta_i, \xi_t) = \mathcal{X}$.

(iv). $\mathcal{X}^2$ *contains at least* $x^{(2)}$ *and* $x^{(2)'}$ *with* $x_j^{(2)} \neq x_j^{(2)'}$ *for all* $j = 1, ..., J$. *Moreover, for all* $x^{(2)} \in \mathcal{X}^2$ *and* $i \in \mathbf{N}$, *there exist* $x^{(1)}, x^{(1')} \in \mathcal{X}^1$, $x_j^{(1)} \neq x_j^{(1')}$ *for all* $j = 1, ..., J$, *such that*

$$(z_i(x^{(1)}; x^{(2)}), x^{(2)}), (z_i(x^{(1')}; x^{(2)}), x^{(2)}) \in \mathcal{X}.$$

(v). *For all* $j = 1, ..., J$, *the level set* $\left\{(z, x^{(2)}) \in \mathcal{Z} : z_j + \beta_{1j}^{(2)} x_j^{(2)} = r_j\right\}$ *is not a singleton for some* $r_j \in \mathbf{R}$.

(vi). *For all* $t \in \mathbf{T}$, $\left\{z + \beta_1^{(2)} x^{(2)} + \xi_t : (z, x^{(2)}) \in \mathcal{Z}\right\} \cap \left\{z + \beta_1^{(2)} x^{(2)} : (z, x^{(2)}) \in \mathcal{Z}\right\} \neq \emptyset$.

Assumption 1′(i) is a multi-index version of Assumption 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 $g(\delta_{it})$ over ξ_t for a given i, satisfies Assumption 1′(i) and is therefore a bijection. Both bijection properties enable to apply the argument of compensating variation as in the single-index case. As argued in Berry et al. (2013), Assumption 1′(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 1′(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 $j = 1, ..., J$ 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 variation but allows for positive cross derivatives in g. Assumptions 1′(ii)-(vi) are similar to those in Assumption 1 and are accommodated to the fact that the compensating vector is of dimension J.

Theorem 1′. *Suppose that Assumptions 1′(i)-(iii) hold.*

- $\beta_i^{(1)}$, α_i, *and* $\beta_i^{(2)} - \beta_1^{(2)}$ *are identified for* $i \in \mathbf{N}$.
- *If Assumptions 1′(iv)-(v) further hold, then*
  - ξ_t *and* $\beta_i^{(2)}$ *are identified for* $i \in \mathbf{N}$ *and* $t \in \mathbf{T}$.
  - $g(v)$ *is identified as a function of indices* $v = \alpha_i + \xi_t + x'\beta_i$.

First, for $i \in \mathbf{N}$ and $x \in \mathcal{X}$, we identify the following vector of quantities using Assumptions 1′(ii)a and (iii):

$$\begin{aligned} &G_i(x) = (\Gamma_{ij}(x))_{j=1}^J, \\ &\Gamma_{ij}(x) = \mathbb{E}\left[\mathbf{1}\left\{y_{it} = j\right\} \mid x_{it} = x, \alpha_i, \beta_i\right] = \int g_j(\alpha_i + x'\beta_i + \xi) dF_\xi(\xi; x^{(2)}) d\xi. \end{aligned} \tag{C.3}$$

To apply the argument of compensating variation 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 injectivity is immediately implied by Assumption 1′(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 1′(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 variation using Assumptions 1′(ii)-(vi) and z_i defined in (C.2). The proof is essentially the same as the single-index case.

D Heterogeneous Slope Across Time

In this section, we given sufficient conditions for the identification of model (3) with heterogeneous slopes across time periods. To start with, define a compensating variable:

$$z^t(x^{(1)}; x^{(2)}) = \xi_t + \beta_t^{(1)} x^{(1)} + x^{(2)}(\beta_t^{(2)} - \beta_1^{(2)}) \tag{D.1}$$

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: $\alpha_i + \xi_1 + \beta_1^{(1)} z^t(x^{(1)}; x^{(2)}) + \beta_1^{(2)} x^{(2)} = \alpha_i + \xi_t + \beta_t^{(1)} x^{(1)} + \beta_t^{(2)} x^{(2)}$. Define $\mathcal{Z}$ as the closure of $\left\{(z^t(x^{(1)}; x^{(2)}), x^{(2)}) : i \in \mathbf{N}, (x^{(1)}, x^{(2)}) \in \mathcal{X}\right\}$.

Assumption 1″.

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

(ii). (a) *For any given* $t \geq 1$*, conditional on* (β_t, ξ_t)*,* $\{(y_{it}, x_{it})\}_{i \geq 2}$ *are independent.*

(b) *For any given* $i \geq 1$*, 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* $(i, i', t) \in \mathbf{N}^2 \times \mathbf{T}$*,* $\alpha_i \perp\!\!\!\perp (\xi_t, \beta_t, x_{it}^{(1)}) \mid x_{it}^{(2)}$*,* $\alpha_i \mid \left\{x_{it}^{(2)} = x^{(2)}\right\} \overset{d}{=} \alpha_{i'} \mid \left\{x_{i't}^{(2)} = x^{(2)}\right\} \sim F_\alpha(\alpha; x^{(2)})$*, and* $Supp(x_{it} \mid \alpha_i, \beta_t, \xi_t) = \mathcal{X}$*.*

(iv). *For all* $x^{(2)} \in \mathcal{X}^2$ *and* $t \in \mathbf{T}$*, there exist* $x^{(1)}, x^{(1')} \in \mathcal{X}^1$*,* $x^{(1)} \neq x^{(1')}$*, such that*

$$\left(z^t(x^{(1)}; x^{(2)}), x^{(2)}\right), \left(z^t(x^{(1')}; x^{(2)}), x^{(2)}\right) \in \mathcal{X}.$$

(v). *The level set* $\left\{(z, x^{(2)}) \in \mathcal{Z} : z + \beta_1^{(2)} x^{(2)} = r\right\}$ *is not a singleton for some* $r \in \mathbf{R}$*.*

(vi). *For all* $i \in \mathbf{N}$*,* $\left\{z + \beta_1^{(2)} x^{(2)} + \alpha_i : (z, x^{(2)}) \in \mathcal{Z}\right\} \cap \left\{z + \beta_1^{(2)} x^{(2)} : (z, x^{(2)}) \in \mathcal{Z}\right\} \neq \emptyset$*.*

Theorem 3. *Suppose that Assumptions 1″(i)-(iv) hold.*

- $\beta_t^{(1)}$*,* ξ_t*, and* $\beta_t^{(2)} - \beta_1^{(2)}$ *are identified for* $t \in \mathbf{T}$*.*
- *If Assumptions 1″(v)-(vi) further hold, then*
  - α_i *and* $\beta_t^{(2)}$ *are identified for* $i \in \mathbf{N}$ *and* $t \in \mathbf{T}$*.*
  - $g(y; v)$ *is identified as a function of* $y \in \mathcal{Y}$ *and index* $v = \alpha_i + \xi_t + x'\beta_t$*.*

Assumption 1″, as well as $z^t(x^{(1)}; x^{(2)})$, mirrors Assumption 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 1 to show Theorem 3. Apart from this difference, the proofs are essentially the same.

E Extension of FPMLE and FPMLE^{++}

E.1 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 $(\beta_1^0, ..., \beta_N^0)'$. We introduce an additional step in FPMLE and FPMLE^{++} to update each slope.
Algorithm FPMLE (Fully Heterogeneous Slopes):

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

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

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

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 $t \in \{1, ..., T\}$:

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

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 $(\alpha_2^{(0)}, ..., \alpha_N^{(0)}, \xi_1^{(0)}, ..., \xi_T^{(0)}, (\boldsymbol{\beta}^{(0)})')' \in \mathcal{A} \times \Xi \times \mathcal{B}^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, ...\}$. 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 $i \in \{1, ..., N\}$:

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

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 $H \subset \{1, ..., K\}$ be the subset indexing variables with heterogeneous coefficients and let $\boldsymbol{\beta}_H^0 \in \mathcal{B}_H^N$ with $\mathcal{B}_H \subset \mathbf{R}^{|H|}$ denote the true heterogeneous parameter values. Let $\beta_{H^c}^0 \in \mathcal{B}_{H^c}$ with $\mathcal{B}_{H^c} \subset \mathbf{R}^{K-|H|}$ denote the true homogeneous parameter values. For any subset $S \subset \{1, ..., K\}$ and vector $u \in \mathbf{R}^K$, let $u_S \in \mathbf{R}^{|S|}$ 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 $(\xi_1^{(0)}, ..., \xi_T^{(0)}, (\beta_{H^c}^{(0)})', (\boldsymbol{\beta}_H^{(0)})')' \in \Xi \times \mathcal{B}_{H^c} \times \mathcal{B}_H^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, ...\}$. Set $s = 0$.

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

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

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 $t \in \{1, ..., T\}$:

$$\xi_t^{(s+1)} \in \arg\max_{\xi \in \Xi_t} \sum_{i=1}^{N} \log g(y_{it}; \alpha_i^{(s+1)} + x'_{H^c,it}\beta_{H^c}^{(s)} + x'_{H,it}\beta_{H,i}^{(s+1)} + \xi).$$

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.[40]

Algorithm FPMLE^{++} (Partly Heterogeneous β_i):

1. Let $(\alpha_2^{(0)}, ..., \alpha_N^{(0)}, \xi_1^{(0)}, ..., \xi_T^{(0)}, (\boldsymbol{\beta}^{(0)})')' \in \mathcal{A} \times \Xi \times \mathcal{B}^N$ be some starting value. Let $\alpha_1^{(j)} = 0$ for all $j \in \{1, 2, ...\}$. 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

3. Compute (keeping it vectorized) for all $i \in \{1, ..., N\}$:

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

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

3. Compute:

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

[40] The "-1" comes from Python's indexing system starting at 0.

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

4. Compute:

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

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.

E.2 Estimating Link Function g

Denote by $\mathcal{G}$ the space of a finite-dimensional parameters, θ_g, that determine the link function g. We abuse the notation $g(\cdot, \theta)$ to refer to the parametrizaton by $\theta^g \in \mathcal{G}$. For instance, when g is the link function corresponding to t-distribution, then θ^g is the degree of freedom of the t-distribution. Another example is the semi-nonparametric estimation of g: for given N and T, $\mathcal{G}$ is a finite-dimensional sieve space of link function g, e.g., $g(\cdot; \theta^g)$ is a linear combination of basis functions $(g_l)_{l=1}^L$ of $\mathcal{G}$, $g(\cdot, \theta) = \sum_{l=1}^L \theta^g_l g_l(\cdot)$.

In these settings, 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 $\mathcal{G}$.

We now show that if FPMLE/FPMLE^{++} converges numerically, then it converges to a stationary point of the likelihood function (6). This property holds straightforwardly for FPMLE as long as the likelihood function is continuously differentiable.

In the next proposition, we prove this property for FPMLE^{++} in the general case with a L-dimensional θ^g being estimated.

Proposition 2. *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 $\mathbf{R}^{N+T+K+L-2}$. 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

$$\theta_i^* = \left[\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right]_{X_i}^+, \quad i = 1, ..., n_{\text{blocs}}. \tag{E.1}$$

Fix $i \in \{1, ..., n_{\text{blocs}}\}$ 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 (E.1) and linearity of the orthogonal projection, it follows that

$$\begin{aligned}
(1/b_i) \times \theta_i^* &= (1/b_i) \times \left[\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right]_{X_i}^+ \\
&= \left[(1/b_i) \times \left(\theta_i^* - \nu \frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*)\right)\right]_{X_i}^+ \\
&= (1/b_i)\theta_i^* - (1/b_i)\frac{\partial \mathscr{L}_{NT}}{\partial \theta_i}(\theta^*).
\end{aligned} \tag{E.2}$$

The result then follows from (E.2) and $(1/b_i) \neq 0$ for all $i = 1, ..., n_{\text{bloc}}$. □

F Consistency in the Presence of Heterogeneous Slopes

Let $\boldsymbol{x} = \{x_{it} : (i,t)\}$, $\boldsymbol{\alpha}^0 = (\alpha_1^0, ..., \alpha_N^0)'$, $\boldsymbol{\xi}^0 = (\xi_1^0, ..., \xi_T^0)'$, $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_N^0)' \in \mathbf{R}^{N \times K}$, $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^0, \boldsymbol{\alpha}^0, \boldsymbol{\xi}^0)$, $\pi_{it}^0 = \alpha_i^0 + \xi_t^0$, $z_{it}^0 = x_{it}'\beta^0 + \pi_{it}^0$, 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

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

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 3.

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

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

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* $k,i,t,N,T,q \le 4$, $\max_{i,t}\mathbb{E}[|\partial_{z^q}\ell_{it}(z_{it}^0)|^{8+\nu}] \le C_1$ *for some* $\nu > 0$, $\mathbb{E}[\exp(\lambda\left|\partial_z\ell_{it}(z_{it}^0)x_{it}^{(k)}\right|)] \le \exp(\lambda C_2)$ *for all* $0 < \lambda < 1/C_2$, *where the expectation* $\mathbb{E}$ *is with respect to* y_{it} *given* z_{it}^0, $\|x_{it}\| \le C_1$, *and* $\inf_{i,T}\sum_{t=1}^{T}(x_{it}^{(k)})^2/T \ge C_3$.

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

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

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* $X_k = (x_{it}^{(k)})_{i=1,...,N;t=1,...,T}$, *and* $\mathbf{1}_n = (1,...,1)' \in \mathbf{R}^n$.

(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* $z \in \mathcal{Z}$, $b_{\min} \le -\partial_{z^2}\ell_{it}(z) \le b_{\max}$ *a.s. uniformly over* i,t,N,T.

Assumption 3 ressembles Assumption 1 in Chen et al. (2021a). The main difference lies on Assumptions 3(iii) and 3(v). Assumption 3(iii) requires the score to have thin tails 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 3(v) adapts the non-collinearity condition to the setting with individual-specific slopes. Along the lines of their non-collinearity condition, it requires the covariates to have sufficient variation once the fixed effects are partialled out. Differently, due to the dimensionality in the number of slope parameters that increases asymptotically (proportionally to N), we instead use the $\|\cdot\|_{\max}$ rather than $\|\cdot\|_{\mathrm{F}}$.[41].

Proposition 3. *Let Assumption 3 hold. Then, as N and T tend to infinity, we have*

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

Proof. We adapt the proof of Lemma 1 in Chen et al. (2021a). For all $z_1, z_2 \in \mathcal{Z}$, a second-order Taylor expansion of $\ell_{it}(z_2)$ around z_1 yields

$$\ell_{it}(z_1) - \ell_{it}(z_2) = [\partial_z \ell_{it}(z_1)](z_1 - z_2) - \frac{1}{2}[\partial_{z^2} \ell_{it}(z^*)](z_1 - z_2)^2, \tag{F.2}$$

for some $z^* \in [\min\{z_1, z_2\}, \max\{z_1, z_2\}]$. Letting $e_{it} = \partial_z \ell_{it} / b_{\min}$, $e := (e_{it})_{i,t}$, we have by definition of $\widehat{\boldsymbol{\theta}}$:

$$\begin{aligned} 0 \geq \mathcal{L}(\boldsymbol{\theta}^0) - \mathcal{L}(\widehat{\boldsymbol{\theta}}) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\ell_{it}(z_{it}^0) - \ell_{it}(\widehat{z}_{it})] \\ &\geq \frac{b_{\min}}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}'(\widehat{\beta}_i - \beta_i^0) + \widehat{\alpha}_i - \alpha_i^0 + \widehat{\xi}_t - \xi_t^0 - e_{it})^2 - \frac{b_{\min}}{2NT}\|e\|_{\mathrm{F}}^2. \end{aligned}$$

Letting $\widehat{\boldsymbol{\lambda}} := (\widehat{\boldsymbol{\alpha}}, \mathbf{1}_N) \in \mathbf{R}^{N\times 2}$, $\widehat{\boldsymbol{f}} := (\mathbf{1}_T, \widehat{\boldsymbol{\xi}}) \in \mathbf{R}^{T\times 2}$, $\boldsymbol{\lambda}^0 := (\boldsymbol{\alpha}^0, \mathbf{1}_N)$, and $\boldsymbol{f}^0 := (\mathbf{1}_T, \boldsymbol{\xi}^0)$, we have

$$\begin{aligned} \frac{b_{\min}}{2NT}\|e\|_{\mathrm{F}}^2 &\geq \frac{b_{\min}}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}'(\widehat{\beta}_i - \beta_i^0) + \widehat{\alpha}_i - \alpha_i^0 + \widehat{\xi}_t - \xi_t^0 - e_{it})^2 \\ &= \frac{b_{\min}}{2NT}\left\|\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\cdot\mathbf{X} + \widehat{\boldsymbol{\lambda}}\widehat{\boldsymbol{f}}' - \boldsymbol{\lambda}^0\boldsymbol{f}^{0\prime} - e\right\|_{\mathrm{F}}^2. \end{aligned}$$

[41] In our setting, $\|\mathbf{v}\|_{\mathrm{F}}^2$ could be of order $O(N)$, while it is $O(1)$ in Chen et al. (2021a). As a result, adopting $\|\mathbf{v}\|_{\mathrm{F}}^2$ in Assumption 3(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.

Because for any matrix A, $\|A\|_{\mathrm{F}}^2 = \mathrm{Tr}\,(AA')$, we 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. (2021a) p.313, we obtain

$$\frac{1}{NT}\,\mathrm{Tr}\,(ee')\geq\frac{1}{NT}\,\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}-e\right)\mathcal{M}_{\widehat{\boldsymbol{f}}}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}-e\right)'\right).\tag{F.3}$$

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 $\mathrm{Tr}(A)\leq\mathrm{rank}(A)\,\|A\|_2$,

$$\begin{aligned}\left|\mathrm{Tr}\left(e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right|&\leq\mathrm{rank}\left(e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\left\|e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right\|_2\leq 2\,\|e\|_2^2,\\|\mathrm{Tr}\,(e'\mathcal{P}_{\boldsymbol{\lambda}^0}e)|&\leq\mathrm{rank}\,(e'\mathcal{P}_{\boldsymbol{\lambda}^0}e)\,\|e'\mathcal{P}_{\boldsymbol{\lambda}^0}e\|_2\leq 2\,\|e\|_2^2,\\\left|\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right|&=\left|\mathrm{Tr}\left(\left[\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}\right]\left[\mathcal{P}_{\boldsymbol{\lambda}^0}e\mathcal{P}_{\widehat{\boldsymbol{f}}}\right]'\right)\right|\leq 2\,\|e\|_2^2.\end{aligned}\tag{F.4}$$

Then, we obtain:

$$\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}e\mathcal{M}_{\widehat{\boldsymbol{f}}}e'\right)\geq\mathrm{Tr}\,(ee')-6\,\|e\|_2^2.\tag{F.5}$$

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 (F.4),

$$\begin{aligned}
\left|\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right)\right| &= \left|\sum_{k=1}^{K}\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]e'\right)\right| \\
&\leq \sum_{k=1}^{K} 2\left\|\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]e'\right\|_2 \\
&\leq 2\left\|e\right\|_2 \times \sum_{k=1}^{K}\left\|\left[\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\mathbf{X}_k\right]\right\|_2 \\
&\leq 2\left\|e\right\|_2 \times \sum_{k=1}^{K}\left\|\mathrm{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)}-\boldsymbol{\beta}^{(k)0}\right)\right\|_2 \times \left\|\mathbf{X}_k\right\|_2 \\
&\leq 2\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}. \\
\left|\mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right| &\leq 2\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}. \\
\left|\mathrm{Tr}\left(\mathcal{P}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{P}_{\widehat{\boldsymbol{f}}}e'\right)\right| &\leq 2\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}.
\end{aligned}$$

Then, we obtain:

$$\mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]\mathcal{M}_{\widehat{\boldsymbol{f}}}e'\right) \geq \mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right) - 6\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\textstyle\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}. \tag{F.6}$$

Plugging (F.5) and (F.6) in (F.3), we obtain:

$$\begin{aligned}
\mathrm{Tr}\left(ee'\right) \geq\ & \mathrm{Tr}\left(ee'\right) + \mathrm{Tr}\left(\mathcal{M}_{\boldsymbol{\lambda}^0}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right)\mathcal{M}_{\widehat{\boldsymbol{f}}}\left(\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right)'\right) \\
& + 2\,\mathrm{Tr}\left(\left[\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right)\cdot\mathbf{X}\right]e'\right) \\
& - 6\left\|e\right\|_2^2 - 12\sqrt{K}\left\|e\right\|_2 \times \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} \times \sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2}.
\end{aligned} \tag{F.7}$$

Under Assumption 3, Lemma S.6 of Fernández-Val and Weidner (2016) holds and implies that $\left\|e\right\|_2 = O_P(N^{5/8})$. Moreover, $\sqrt{\sum_{k=1}^{K}\left\|\mathbf{X}_k\right\|_{\mathrm{F}}^2} = O_P(\sqrt{NT}) = O_P(N)$.

Now, denote by $\tilde{\boldsymbol{\beta}}$ the solutions of the MLE with $\alpha_i = \alpha_i^0$ and $\xi_t = \xi_t^0$ for $i = 1, ..., N$ and $t = 1, ..., T$. We prove the following lemma.

Lemma 2. *Suppose that Assumption 3 holds. Then,* $\left\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_{\max} = O_P\left(\sqrt{\frac{\ln N}{N}}\right)$.

Proof. Note that the first-order condition of the MLE with respect to $\tilde{\boldsymbol{\beta}}$: for

$k = 1, ..., K$ and $i = 1, ..., N$,

$$\sum_{t=1}^{T} \partial_z \ell_{it}(\tilde{z}_{it}) x_{it}^{(k)} = 0. \tag{F.8}$$

Then, using the first-order Taylor expansion at $\beta_i^{(k)0}$, we obtain:

$$\begin{aligned} 0 &= \sum_{t=1}^{T} \partial_z \ell_{it}(z_{it}^0) x_{it}^{(k)} + \sum_{t=1}^{T} \partial_{z^2} \ell_{it}(z_{it}^*)(x_{it}^{(k)})^2 \times (\tilde{\beta}_i^{(k)} - \beta_i^{(k)0}) \\ \implies \tilde{\beta}_i^{(k)} - \beta_i^{(k)0} &= \frac{\frac{1}{T}\sum_{t=1}^{T} \partial_z \ell_{it}(z_{it}^0) x_{it}^{(k)}}{\frac{1}{T}\sum_{t=1}^{T} -\partial_{z^2} \ell_{it}(z_{it}^*)(x_{it}^{(k)})^2}, \end{aligned}$$

where z_{it}^* is between z_{it}^0 and $\tilde{z}_{it}$. Since Assumption 3(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 $\{x_{it}\}_{i=1,...,N;t=1,...,T}, \boldsymbol{\alpha}^0, \boldsymbol{\beta}^0, \boldsymbol{\xi}^0$, $S_{i,T}$ 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 3(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:

$$\begin{aligned} \Pr\left(\left\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_{\max} \leq A\sqrt{\frac{\ln N}{N}}\right) &= \prod_{i=1}^{N} \Pr\left(\left|\tilde{\beta}_i - \beta_i^0\right| \leq A\sqrt{\frac{\ln N}{N}}\right) \\ &\geq \prod_{i=1}^{N} \Pr\left(|S_{i,T}| \leq A b_{\min} C_2 \sqrt{\frac{\ln N}{N}}\right) \\ &= \prod_{i=1}^{N} \left(1 - \Pr\left(|S_{i,T}| > A b_{\min} C_2 \sqrt{\frac{\ln N}{N}}\right)\right) \\ &\geq \left(1 - 2\exp\left\{-A^2 B \times \ln N\right\}\right)^N \\ &= \left(1 - \frac{2}{N^{A^2 B}}\right)^N, \end{aligned}$$

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:

$$\Pr\left(\left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \leq A \sqrt{\frac{\ln N}{N}} \right) \to 1.$$

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 (F.8), we obtain that: for $k = 1, ..., K$ and $i = 1, ..., N$,

$$\begin{aligned} \sum_{t=1}^{T} \partial_z \ell_{it}(\tilde{z}_{it}) x_{it}^{(k)} = 0 \implies & \text{ The diagonal elements of } \mathbf{X}_k \tilde{e}' \text{ are zeros.} \\ \implies & \operatorname{Tr}\left(\operatorname{Diag}\left(\widehat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)0} \right) \mathbf{X}_k \tilde{e}' \right) = 0, \end{aligned}$$

where $\tilde{e} = (\partial_z \ell_{it}(x_{it}' \tilde{\beta}_i + \alpha_i^0 + \xi_t^0))_{i=1,...,N;t=1,...,T}$. Then,

$$\begin{aligned} \left| \operatorname{Tr}\left(\left[\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right) \cdot \mathbf{X} \right] e' \right) \right| &= \left| \operatorname{Tr}\left(\sum_{k=1}^{K} \left[\operatorname{Diag}\left(\tilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)0} \right) \mathbf{X}_k \right] (e - \tilde{e})' \right) \right| \\ &= \left| \operatorname{Tr}\left(\sum_{k=1}^{K} \left[\operatorname{Diag}\left(\tilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)0} \right) \mathbf{X}_k \right] \left[\int_0^1 \partial_{z^2} \ell_{it} \left(x_{it}'(t\beta_i^0 + (1-t)\tilde{\beta}_i) + \alpha_i^0 + \xi_t^0 \right) \sum_{k=1}^{K} (\beta_i^{(k)0} - \tilde{\beta}_i^{(k)}) x_{it}^{(k)} dt \right]'_{i=1,...,N;t=1,...,T} \right) \right| \\ &\leq b_{\max} \operatorname{Tr}\left(\sum_{k=1}^{K} \left[\operatorname{Diag}\left(\left| \tilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)0} \right| \right) |\mathbf{X}_k| \right] \sum_{k=1}^{K} \left[\operatorname{Diag}\left(\left| \tilde{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}^{(k)0} \right| \right) |\mathbf{X}_k| \right]' \right) \\ &\leq K b_{\max} \left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \times \left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \times \sum_{k=1}^{K} \|\mathbf{X}_k\|_{\mathrm{F}}^2 \end{aligned} \tag{F.9}$$

Plugging (F.9) in (F.7), we obtain:

$$\begin{aligned} & 6 \|e\|_2^2 + 2 \sqrt{K \sum_{k=1}^{K} \|\mathbf{X}_k\|_{\mathrm{F}}^2} \left(6 \|e\|_2 + b_{\max} \left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \sqrt{K \sum_{k=1}^{K} \|\mathbf{X}_k\|_{\mathrm{F}}^2} \right) \times \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \\ & \geq \operatorname{Tr}\left(\mathcal{M}_{(\boldsymbol{\alpha}^0, \mathbf{1}_N)} \left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right) \cdot \mathbf{X} \right) \mathcal{M}_{(\mathbf{1}_T, \widehat{\boldsymbol{\xi}})} \left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right) \cdot \mathbf{X} \right)' \right). \end{aligned}$$

Using Assumption 3(v), we obtain that

$$\begin{aligned} & \frac{1}{NT} \left[6 \|e\|_2^2 + 2 \sqrt{K \sum_{k=1}^{K} \|\mathbf{X}_k\|_{\mathrm{F}}^2} \left(6 \|e\|_2 + b_{\max} \left\| \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \sqrt{K \sum_{k=1}^{K} \|\mathbf{X}_k\|_{\mathrm{F}}^2} \right) \times \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max} \right] \\ & \geq c \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_{\max}^2. \end{aligned}$$

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

Proposition 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

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 (F.1), $\mathbf{B}_i$, is a subset of $[-\ln N, \ln N]$ for $i \leq N$, and $\Pr(\beta_i^0 \in \mathbf{B}_i, i = 1, ..., N|\mathbf{X}_N) \to 1$.[42] 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).

G Monte Carlo Experiments: Additional 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:

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

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

$$\widehat{\delta}_{NT} = \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T \widehat{\beta}_{NT}\Lambda'(x_{it}\widehat{\beta}_{NT} + \widehat{\alpha}_{NT,i} + \widehat{\xi}_{NT,t}). \tag{G.2}$$

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

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 a Hessian stepsize and 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 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 FPMLE^{++}, we use a step size of $\nu^{(s)} \approx 1/(NT)$, or an Hessian step. 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.

G.1 Additional Tables

Table G.1: Numerical Convergence, $N = 5000$, $T = 30$

	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 G.2: Numerical Convergence, $N = T = 200$

		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.

G.2 Poisson Count Model with Heterogeneous Slopes

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

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

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

Table G.3 summarizes the bias of $\widehat{\beta}_i$ (estimated by FPMLE and FPMLE^{++}) and its numerical convergence to the MLE estimator.

Table G.3: 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 $b \in \{1, ..., B\}$:

Table G.4: Inference – Poisson Model with Heterogeneous Slopes

	$\widehat{\mathbb{E}}(\beta_{0i})$			$\sqrt{\widehat{\text{Var}}(\beta_{0i})}$			$\widehat{\text{APE}}$		
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 G.2 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.

(a) Draw N^* units from $\{1, ..., N\}$ 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 $\mathcal{I}^{(b)}$:

$$\left\{\widehat{\beta}_{i,\text{fs}}^{(b)} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\alpha}_{i,\text{fs}}^{(b)} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\xi}_{t,\text{fs}}^{(b)} : t = 1, ..., T\right\}.$$

(c) Compute half-sample FPMLE^{++} estimates using only units in $\mathcal{I}_{:\lfloor N^*/2 \rfloor}^{(b)}$:

$$\left\{\widehat{\beta}_{i,1\text{N}^*}^{(b)} : i \in \mathcal{I}_{:\lfloor N^*/2 \rfloor}^{(b)}\right\}, \quad \left\{\widehat{\alpha}_{i,1\text{N}^*}^{(b)} : i \in \mathcal{I}_{:\lfloor N^*/2 \rfloor}^{(b)}\right\}, \quad \left\{\widehat{\xi}_{t,1\text{N}^*}^{(b)} : t = 1, ..., T\right\}.$$

Compute half-sample FPMLE^{++} estimates using only units in $\mathcal{I}_{\lfloor N^*/2 \rfloor:}^{(b)}$:

$$\left\{\widehat{\beta}_{i,2\text{N}^*}^{(b)} : i \in \mathcal{I}_{\lfloor N^*/2 \rfloor:}^{(b)}\right\}, \quad \left\{\widehat{\alpha}_{i,2\text{N}^*}^{(b)} : i \in \mathcal{I}_{\lfloor N^*/2 \rfloor:}^{(b)}\right\}, \quad \left\{\widehat{\xi}_{t,2\text{N}^*}^{(b)} : t = 1, ..., T\right\}.$$

(d) Compute half-sample FPMLE^{++} estimates using only time periods in $\{1, ..., \lfloor T/2 \rfloor\}$:

$$\left\{\widehat{\beta}_{i,1\text{T}}^{(b)} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\alpha}_{i,1\text{T}}^{(b)} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\xi}_{t,1\text{T}}^{(b)} : t = 1, ..., \lfloor T/2 \rfloor\right\}.$$

Compute half-sample FPMLE^{++} estimates using only time periods in $\{\lfloor T/2 \rfloor + 1, ..., T\}$:

$$\left\{\widehat{\beta}_{i,2\text{T}}^{(b)} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\alpha}_{i,2\text{T}}^{(b)} : i \in \mathcal{I}^{(b)}\right\}, \quad \left\{\widehat{\xi}_{t,2\text{T}}^{(b)} : t = \lfloor T/2 \rfloor + 1, ..., T\right\}.$$

(e) Let

$$\begin{aligned}
\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) \\
&\quad - \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} \\
&\quad - \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. \\
&\quad \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) \\
&\quad - \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. \\
&\quad \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) \\
&\quad - \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. \\
&\quad \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) \\
&\quad - \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. \\
&\quad \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),
\end{aligned}$$

and $\widehat{\mu} := (\widehat{\mu}^{(1)}, ..., \widehat{\mu}^{(B)})$, $\widehat{\sigma} := (\widehat{\sigma}^{(1)}, ..., \widehat{\sigma}^{(B)})$, and $\widehat{\mathrm{APE}} := (\widehat{\mathrm{APE}}^{(1)}, ..., \widehat{\mathrm{APE}}^{(B)})$.

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 $\mathbf{X}$.

H Empirical Illustrations: Additional Results

Table H.5: 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 H.6: 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 H.7: 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 (14) and (15). The observed characteristics z_i are the same as in Table H.6 .