Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Individual and time effects in nonlinear panel models with large N, T

Iván Fernández-Val [a,*], Martin Weidner [b,c]

[a] Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215-1403, USA
[b] Department of Economics, University College London, Gower Street, London WC1E 6BT, UK
[c] Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 7 Ridgmount Street, London WC1E 7AE, UK

ABSTRACT

We derive fixed effects estimators of parameters and average partial effects in (possibly dynamic) nonlinear panel data models with individual and time effects. They cover logit, probit, ordered probit, Poisson and Tobit models that are important for many empirical applications in micro and macroeconomics. Our estimators use analytical and jackknife bias corrections to deal with the incidental parameter problem, and are asymptotically unbiased under asymptotic sequences where N/T converges to a constant. We develop inference methods and show that they perform well in numerical examples.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fixed effects estimators of nonlinear panel data models can be severely biased because of the incidental parameter problem (Neyman and Scott, 1948). A growing literature, surveyed in Arellano and Hahn (2007), shows that the leading term of an asymptotic expansion of the bias as both the cross-sectional dimension N and time series dimension T of the panel grow, can be characterized and corrected. In models with individual effects, the leading bias term is of order $1/T$ and comes from the estimation of the individual effects. This result, however, does not apply to models with individual and time effects, where both of these effects are treated as parameters to be estimated. In this paper we show that the estimation of the time effects causes an additional incidental parameter bias of order $1/N$. Thus, if N and T are similarly large, the bias produced by the estimation of the time effects is of similar order of magnitude to the bias produced by the estimation of the individual effects, and both biases need to be corrected. We provide the corresponding analytical and jackknife bias corrections.

The asymptotic approximation to the fixed effects estimators that lets the two dimensions of the panel to grow with the sample size is motivated by the recent availability of long panels and other large pseudo-panel data structures where the indices might not correspond to individuals and time periods. Examples of these data sets include traditional microeconomic panel surveys with a long history of data such as the PSID and NLSY, international cross-country panels such as the Penn World Table, U.S. state level panels over time such as the CPS, and square pseudo-panels of trade flows across countries such as the Feenstra's World Trade Flows and CEPII, where the indices correspond to the same countries indexed as importers and exporters.

* Corresponding author.
E-mail addresses: ivanf@bu.edu (I. Fernández-Val), m.weidner@ucl.ac.uk (M. Weidner).

We focus on semi-parametric models with log-likelihood functions that are concave in all parameters, and where each individual effect α_i and time effect γ_t enter the log-likelihood for observation (i, t) additively as $\alpha_i + \gamma_t$. This is the most common specification for the individual and time effects in linear models and is also a natural specification in the nonlinear models that we consider. Imposing concavity of the log-likelihood function greatly facilitates showing consistency in our setting where the dimension of the parameter space grows with the sample size. The most popular limited dependent variable models, including logit, probit, ordered probit, Tobit and Poisson models have concave log-likelihood functions, possibly after reparametrization (Olsen, 1978, and Pratt, 1981). We note here that the general expansion that we derive in Appendix B do not impose additivity and concavity, but we use these restrictions to apply the expansion to fixed effects estimators. The models that we consider are semi-parametric because the joint distribution of the explanatory variables and the unobserved effects is left unspecified. The explanatory variables can be either strictly exogenous or predetermined.

We derive bias expansions and corrections for fixed effects estimators of common parameters β and average partial effects (APEs). The vector β includes all the unknown parameters that enter the log-likelihood function other than the individual and time effects, such as index coefficients in a probit model. The APEs are functions of the data, the common parameters, and the individual and time effects in nonlinear models. We find that the properties of the fixed effects estimators of β and the APEs are different. For β, the order of the bias is $1/T + 1/N$, which is of the same as the rate of convergence $1/\sqrt{NT}$ under sequences where N/T converges to a constant. For the APEs, we uncover that the incidental parameter problem is negligible asymptotically because the order of the bias, $1/N + 1/T$, is smaller than the rate of convergence, which is $1/\sqrt{N} + 1/\sqrt{T}$, slower than for model parameters. To the best of our knowledge, this rate result is new for fixed effects estimators of average partial effects in nonlinear panel models with individual and time effects.[1] In numerical examples we find that the bias corrections, while not necessary to center the asymptotic distribution of APE estimators, do improve their finite-sample properties, specially in dynamic models.

The bias correction eliminates the bias terms of orders $1/T$ and $1/N$ from the fixed effects estimators. We consider two methods to implement the correction: an analytical bias correction similar to Hahn and Newey (2004) and Hahn and Kuersteiner (2011), and a suitable modification of the split panel jackknife of Dhaene and Jochmans (2015).[2] However, the theory of the previous papers does not cover the models that we consider, because, in addition to not allowing for time effects, it assumes either identical distribution or stationarity over time for the processes of the observed variables, conditional on the unobserved effects. These assumptions are violated in our models due to the presence of the time effects, so we need to adjust the asymptotic theory accordingly. The individual and time effects introduce strong correlation in both dimensions of the panel. Conditional on the unobserved effects, we impose cross-sectional independence and weak time-serial dependence, and we allow for heterogeneity in both dimensions.

Simulation evidence indicates that our corrections improve the estimation and inference performance of the fixed effects estimators of parameters and average effects. The analytical corrections dominate the jackknife corrections in a probit model for sample sizes that are relevant for empirical practice. In the online supplement (see Appendix E), we illustrate the corrections with an empirical application on the relationship between competition and innovation using a panel of U.K. industries, following Aghion et al. (2005). We find that the inverted-U pattern relationship found by Aghion et al. is robust to relaxing the strict exogeneity assumption of competition with respect to the innovation process and to the inclusion of innovation dynamics. We also uncover substantial positive state dependence in the innovation process.

Literature review. The Neyman and Scott incidental parameter problem has been extensively discussed in the econometric literature; see, for example, Heckman (1981), Lancaster (2000), and Greene (2004). There is also a vast literature that shows how to tackle the problem in specific models under asymptotic sequences where T is fixed and N grows to infinity. However, there are results, e.g. from Honoré and Tamer (2006), Chamberlain (2010), and Chernozhukov et al. (2013), showing that model parameters and APEs are not point identified in important nonlinear panel data models under fixed-T asymptotic sequences, implying that no fixed-T consistent point estimators exist in these models.

A recent response to the incidental parameter problem is to adopt an alternative asymptotic approximation where both N and T grow with the sample size. Under these large-T sequences, the fixed effects estimator is consistent but has bias in the asymptotic distribution. This asymptotic bias is the large-T version of the incidental parameter problem and has motivated the development of bias corrections. Examples of papers that use this approximation include Phillips and Moon (1999), Hahn and Kuersteiner (2002), Lancaster (2002), Woutersen (2002), Alvarez and Arellano (2003), Hahn and Newey (2004), Carro (2007), Arellano and Bonhomme (2009), Fernández-Val (2009), Hahn and Kuersteiner (2011), Fernández-Val and Vella (2011) and Kato et al. (2012). This previous work, however, does not cover models with time effects.[3] Our contribution to this literature is to extend the large-T bias corrections to models with two-way unobserved effects such as the individual and time effects commonly included in linear models.

The large-T panel literature on models with both individual and time effects is sparse. Pesaran (2006), Bai (2009) and Moon and Weidner (forthcoming, 2015b) study linear regression models with interactive individual and time fixed effects. The fixed effects estimators in these models also have asymptotic bias of order $1/T + 1/N$, but the methods used to derive this bias rely on linearity and therefore cannot be applied to the nonlinear models that we consider. Hahn and Moon (2006) consider bias corrected fixed effects estimators in panel linear autoregressive models with additive individual and time effects. Regarding non-linear models, there is independent and contemporaneous work by Charbonneau (2012, 2014), which extends the conditional fixed effects estimators to logit and Poisson models with individual and time effects. She differences out the individual and time effects by conditioning on sufficient statistics. The conditional approach completely eliminates the asymptotic bias coming from the estimation of the incidental parameters, but it does not permit estimation of average partial effects and has not been developed for models with predetermined regressors. We instead consider estimators of model parameters and average partial effects in nonlinear models with predetermined regressors. The two approaches can therefore be considered as complementary.

---

[1] Galvao and Kato (2014) also found slow rates of convergence for fixed effects estimators in linear models with individual effects under misspecification. Fernández-Val and Lee (2013) pointed out this issue in nonlinear models with only individual effects.

[2] A similar split panel jackknife bias correction method was outlined in Hu (2002).

---

[3] An exception is Woutersen (2002), which considers a special type of grouped time effects whose number is fixed with T. We instead consider an unrestricted set of T time effects, one for each time period.

Outline of the paper. The rest of the paper is organized as follows. Section 2 introduces the model and fixed effects estimators. Section 3 describes the bias corrections to deal with the incidental parameter problem and illustrates how the bias corrections work through an example. Section 4 provides the asymptotic theory. Section 5 presents Monte Carlo results. The Appendix collects the proofs of the main results, and an online supplement to the paper contains additional technical derivations, numerical examples, and an empirical application (Fernández-Val and Weidner, 2015b) (see Appendix E).

2. Model and estimators

2.1. Model

The data consist of $N \times T$ observations $\{(Y_{it}, X_{it}')' : 1 \leq i \leq N, 1 \leq t \leq T\}$, for a scalar outcome variable of interest Y_{it} and a vector of explanatory variables X_{it}. We assume that the outcome for individual i at time t is generated by the sequential process:

$$Y_{it} \mid X_i^t, \alpha, \gamma, \beta \sim f_Y(\cdot \mid X_{it}, \alpha_i, \gamma_t, \beta),$$
$$(i = 1, \ldots, N; t = 1, \ldots, T),$$

where $X_i^t = (X_{i1}, \ldots, X_{it})$, $\alpha = (\alpha_1, \ldots, \alpha_N)$, $\gamma = (\gamma_1, \ldots, \gamma_T)$, f_Y is a known probability function, and β is a finite dimensional parameter vector. The variables α_i and γ_t are unobserved individual and time effects that in economic applications capture individual heterogeneity and aggregate shocks, respectively. The model is semiparametric because we do not specify the distribution of these effects nor their relationship with the explanatory variables. The conditional distribution f_Y represents the parametric part of the model. The vector X_{it} contains predetermined variables with respect to Y_{it}. Note that X_{it} can include lags of Y_{it} to accommodate dynamic models.

We consider two running examples throughout the analysis:

Example 1 (*Binary Response Model*). Let Y_{it} be a binary outcome and F be a cumulative distribution function, e.g. the standard normal or standard logistic distribution. We can model the conditional distribution of Y_{it} using the single-index specification with individual and time effects

$$f_Y(y \mid X_{it}, \alpha_i, \gamma_t, \beta) = F(X_{it}'\beta + \alpha_i + \gamma_t)^y$$
$$\times [1 - F(X_{it}'\beta + \alpha_i + \gamma_t)]^{1-y}, \quad y \in \{0, 1\}.$$

In a labor economics application, Y can be an indicator for female labor force participation and X can include fertility indicators and other socio-economic characteristics.

Example 2 (*Poisson Model*). Let Y_{it} be a non-negative integer-valued outcome, and $f(\cdot; \lambda)$ be the probability mass function of a Poisson random variable with mean $\lambda > 0$. We can model the conditional distribution of Y_{it} using the single index specification with individual and time effects

$$f_Y(y \mid X_{it}, \alpha_i, \gamma_t, \beta) = f(y; \exp[X_{it}'\beta + \alpha_i + \gamma_t]),$$
$$y \in \{0, 1, 2, \ldots.\}.$$

In an industrial organization application, Y can be the number of patents that a firm produces and X can include investment in R&D and other firm characteristics.

For estimation, we adopt a fixed effects approach, treating the realization of the unobserved individual and time effects as parameters to be estimated. We collect all these effects in the vector $\phi_{NT} = (\alpha_1, \ldots, \alpha_N, \gamma_1, \ldots, \gamma_T)'$. The model parameter β usually includes regression coefficients of interest, while the vector ϕ_{NT} is treated as a nuisance parameter. The true values of the parameters, denoted by β^0 and $\phi_{NT}^0 = (\alpha_1^0, \ldots, \alpha_N^0, \gamma_1^0, \ldots, \gamma_T^0)'$, are the solution to the population conditional maximum likelihood problem

$$\max_{(\beta, \phi_{NT}) \in \mathbb{R}^{\dim \beta + \dim \phi_{NT}}} \mathbb{E}_\phi[\mathcal{L}_{NT}(\beta, \phi_{NT})],$$

$$\mathcal{L}_{NT}(\beta, \phi_{NT}) := (NT)^{-1/2} \left\{ \sum_{i,t} \log f_Y(Y_{it} \mid X_{it}, \alpha_i, \gamma_t, \beta) \right.$$

$$\left. - b(v_{NT}'\phi_{NT})^2/2 \right\}, \tag{2.1}$$

for every N, T, where $\mathbb{E}_\phi$ denotes the expectation with respect to the distribution of the data conditional on the unobserved effects and initial conditions including strictly exogenous variables, $b > 0$ is an arbitrary constant, $v_{NT} = (1_N', -1_T')'$, and 1_N and 1_T denote vectors of ones with dimensions N and T. Existence and uniqueness of the solution to the population problem will be guaranteed by our assumptions in Section 4, including concavity of the objective function in all parameters. The second term of $\mathcal{L}_{NT}$ is a penalty that imposes a normalization needed to identify ϕ_{NT} in models with scalar individual and time effects that enter additively into the log-likelihood function as $\alpha_i + \gamma_t$.[4] In this case, adding a constant to all α_i, while subtracting it from all γ_t, does not change $\alpha_i + \gamma_t$. To eliminate this ambiguity, we normalize ϕ_{NT}^0 to satisfy $v_{NT}'\phi_{NT}^0 = 0$, i.e. $\sum_i \alpha_i^0 = \sum_t \gamma_t^0$. The penalty produces a maximizer of $\mathcal{L}_{NT}$ that is automatically normalized. We could equivalently impose $v_{NT}'\phi_{NT} = 0$ as a constraint, but for technical reasons we prefer to work with an unconstrained optimization problem. There are other possible normalizations for ϕ_{NT}, such as $\alpha_1 = 0$. The model parameter β is invariant to the choice of normalization, that is, our asymptotic results on the estimator for β are independent of this choice of normalization. Our choice is convenient for certain intermediate results that involve the incidental parameter ϕ_{NT}, its score vector and its Hessian matrix. The pre-factor $(NT)^{-1/2}$ in $\mathcal{L}_{NT}(\beta, \phi_{NT})$ is just a rescaling.

Other quantities of interest involve averages over the data and unobserved effects

$$\delta_{NT}^0 = \mathbb{E}[\Delta_{NT}(\beta^0, \phi_{NT}^0)],$$
$$\Delta_{NT}(\beta, \phi_{NT}) = (NT)^{-1} \sum_{i,t} \Delta(X_{it}, \beta, \alpha_i, \gamma_t), \tag{2.2}$$

where $\mathbb{E}$ denotes the expectation with respect to the joint distribution of the data and the unobserved effects, provided that the expectation exists. δ_{NT}^0 is indexed by N and T because the marginal distribution of $\{(X_{it}, \alpha_i, \gamma_t) : 1 \leq i \leq N, 1 \leq t \leq T\}$ can be heterogeneous across i and/or t; see Section 4.2. These averages include average partial effects (APEs), which are often the ultimate quantities of interest in nonlinear models. The APEs are invariant to the choice of normalization for ϕ_{NT} if α_i and γ_t enter $\Delta(X_{it}, \beta, \alpha_i, \gamma_t)$ as $\alpha_i + \gamma_t$. Some examples of partial effects that satisfy this condition are the following:

Example 1 (*Binary Response Model*). If $X_{it,k}$, the kth element of X_{it}, is binary, its partial effect on the conditional probability of Y_{it} is

$$\Delta(X_{it}, \beta, \alpha_i, \gamma_t) = F(\beta_k + X_{it,-k}'\beta_{-k} + \alpha_i + \gamma_t)$$

---

[4] In Appendix B we derive asymptotic expansions that apply to general models with multiple unobserved effects. In order to use these expansions to obtain the asymptotic distribution of the panel fixed effects estimators, we need to derive the properties of the expected Hessian of the incidental parameters, a matrix with increasing dimension, and to show the consistency of the estimator of the incidental parameter vector. The additive specification $\alpha_i + \gamma_t$ is useful to characterize the Hessian and we impose strict concavity of the objective function to show the consistency.

$$-F(X'_{it,-k}\beta_{-k} + \alpha_i + \gamma_t), \qquad (2.3)$$

where β_k is the kth element of β, and $X_{it,-k}$ and β_{-k} include all elements of X_{it} and β except for the kth element. If $X_{it,k}$ is continuous and F is differentiable, the partial effect of $X_{it,k}$ on the conditional probability of Y_{it} is

$$\Delta(X_{it}, \beta, \alpha_i, \gamma_t) = \beta_k \partial F(X'_{it}\beta + \alpha_i + \gamma_t), \qquad (2.4)$$

where ∂F is the derivative of F.

Example 2 (*Poisson Model*)**.** If X_{it} includes Z_{it} and some known transformation $H(Z_{it})$ with coefficients β_k and β_j, the partial effect of Z_{it} on the conditional expectation of Y_{it} is

$$\Delta(X_{it}, \beta, \alpha_i, \gamma_t) = [\beta_k + \beta_j \partial H(Z_{it})] \exp(X'_{it}\beta + \alpha_i + \gamma_t). \qquad (2.5)$$

2.2. Fixed effects estimators

We estimate the parameters by solving the sample analog of problem (2.1), i.e.

$$\max_{(\beta, \phi_{NT}) \in \mathbb{R}^{\dim \beta + \dim \phi_{NT}}} \mathcal{L}_{NT}(\beta, \phi_{NT}). \qquad (2.6)$$

As in the population case, we shall impose conditions guaranteeing that the solution to this maximization problem exists and is unique with probability approaching one as N and T become large. For computational purposes, we note that the solution to the program (2.6) for β is the same as the solution to the program that imposes $v'_{NT}\phi_{NT} = 0$ directly as a constraint in the optimization, and is invariant to the normalization. In our numerical examples we impose either $\alpha_1 = 0$ or $\gamma_1 = 0$ directly by dropping the first individual or time effect. This constrained program has good computational properties because its objective function is concave and smooth in all the parameters. We have developed the commands `probitfe` and `logitfe` in `Stata` to implement the methods of the paper for probit and logit models (Cruz-González et al., 2015).[5] When N and T are large, e.g., $N > 2{,}000$ and $T > 50$, we recommend the use of optimization routines that exploit the sparsity of the design matrix of the model to speed up computation such as the package `Speedglm` in R (Enea, 2012). For a probit model with $N = 2{,}000$ and $T = 52$, `Speedglm` computes the fixed effects estimator in less than 2 min with a 2×2.66 GHz 6-Core Intel Xeon processor, more than 7.5 times faster than our `Stata` command `probitfe` and more than 30 times faster than the R command `glm`.[6]

To analyze the statistical properties of the estimator of β it is convenient to first concentrate out the nuisance parameter ϕ_{NT}. For given β, we define the optimal $\widehat{\phi}_{NT}(\beta)$ as

$$\widehat{\phi}_{NT}(\beta) = \underset{\phi_{NT} \in \mathbb{R}^{\dim \phi_{NT}}}{\operatorname{argmax}} \mathcal{L}_{NT}(\beta, \phi_{NT}). \qquad (2.7)$$

The fixed effects estimators of β and ϕ_{NT} are

$$\widehat{\beta}_{NT} = \underset{\beta \in \mathbb{R}^{\dim \beta}}{\operatorname{argmax}} \mathcal{L}_{NT}(\beta, \widehat{\phi}_{NT}(\beta)), \qquad \widehat{\phi}_{NT} = \widehat{\phi}_{NT}(\widehat{\beta}). \qquad (2.8)$$

Estimators of APEs can be formed by plugging-in the estimators of the model parameters in the sample version of (2.2), i.e.

$$\widehat{\delta}_{NT} = \Delta_{NT}(\widehat{\beta}, \widehat{\phi}_{NT}). \qquad (2.9)$$

Again, $\widehat{\delta}_{NT}$ is invariant to the normalization chosen for ϕ_{NT} if α_i and γ_t enter $\Delta(X_{it}, \beta, \alpha_i, \gamma_t)$ as $\alpha_i + \gamma_t$.

---

[5] We refer to this companion work for computational details.

[6] Additional comparisons of computational times are available from the authors upon request.

3. Incidental parameter problem and bias corrections

In this section we give a heuristic discussion of the main results, leaving the technical details to Section 4. We illustrate the analysis with numerical calculations based on a variation of the classical Neyman and Scott (1948) variance example.

3.1. Incidental parameter problem

Fixed effects estimators in nonlinear models suffer from the incidental parameter problem (Neyman and Scott, 1948). The source of the problem is that the dimension of the nuisance parameter ϕ_{NT} increases with the sample size under asymptotic approximations where either N or T passes to infinity. To describe the problem let

$$\overline{\beta}_{NT} := \underset{\beta \in \mathbb{R}^{\dim \beta}}{\operatorname{argmax}} \mathbb{E}_\phi \left[\mathcal{L}_{NT}(\beta, \widehat{\phi}_{NT}(\beta)) \right]. \qquad (3.1)$$

The fixed effects estimator is inconsistent under the traditional Neyman and Scott asymptotic sequences where $N \to \infty$ and T is fixed, i.e., $\text{plim}_{N\to\infty} \overline{\beta}_{NT} \neq \beta^0$. Similarly, the fixed effects estimator is inconsistent under asymptotic sequences where $T \to \infty$ and N is fixed, i.e., $\text{plim}_{T\to\infty} \overline{\beta}_{NT} \neq \beta^0$. Note that $\overline{\beta}_{NT} = \beta^0$ if $\widehat{\phi}_{NT}(\beta)$ is replaced by $\phi_{NT}(\beta) = \operatorname{argmax}_{\phi_{NT} \in \mathbb{R}^{\dim \phi_{NT}}} \mathbb{E}_\phi[\mathcal{L}_{NT}(\beta, \phi_{NT})]$. Under asymptotic approximations where either N or T is fixed, there is only a fixed number of observations to estimate some of the components of ϕ_{NT}, T for each individual effect or N for each time effect, rendering the estimator $\widehat{\phi}_{NT}(\beta)$ inconsistent for $\phi_{NT}(\beta)$. The nonlinearity of the model propagates the inconsistency to the estimator of β.

A key insight of the large-T panel data literature is that the incidental parameter problem becomes an asymptotic bias problem under an asymptotic approximation where $N \to \infty$ and $T \to \infty$ (e.g., Arellano and Hahn, 2007). For models with only individual effects, this literature derived the expansion $\overline{\beta}_{NT} = \beta^0 + B/T + o_P(T^{-1})$ as $N, T \to \infty$, for some constant B. The fixed effects estimator is consistent because $\text{plim}_{N,T\to\infty} \overline{\beta}_{NT} = \beta^0$, but has bias in the asymptotic distribution if B/T is not negligible relative to $1/\sqrt{NT}$, the order of the standard deviation of the estimator. This asymptotic bias problem, however, is easier to tackle than the inconsistency problem that arises under the traditional Neyman and Scott asymptotic approximation. We show that the same insight still applies to models with individual and time effects, but with a different expansion for $\overline{\beta}_{NT}$. We characterize the expansion and develop bias corrections.

3.2. Bias expansions and bias corrections

Some expansions can be used to explain our corrections. For smooth likelihoods and under appropriate regularity conditions, as $N, T \to \infty$,

$$\overline{\beta}_{NT} = \beta^0 + \overline{B}_\infty^\beta/T + \overline{D}_\infty^\beta/N + o_P(T^{-1} \vee N^{-1}), \qquad (3.2)$$

for some $\overline{B}_\infty^\beta$ and $\overline{D}_\infty^\beta$ that we characterize in Theorem 4.1 and explain in Remark 2, where $a \vee b := \max(a, b)$. Unlike in nonlinear models without incidental parameters, the order of the bias is higher than the inverse of the sample size $(NT)^{-1}$ due to the slow rate of convergence of $\widehat{\phi}_{NT}$. Note also that by the properties of the maximum likelihood estimator

$$\sqrt{NT}(\widehat{\beta}_{NT} - \overline{\beta}_{NT}) \to_d \mathcal{N}(0, \overline{V}_\infty),$$

for some $\overline{V}_\infty$ that we also characterize in Theorem 4.1. Under asymptotic sequences where $N/T \to \kappa^2$ as $N, T \to \infty$, the fixed

effects estimator is asymptotically biased because

$$
\begin{aligned}
\sqrt{NT}(\widehat{\beta}_{NT} - \beta^0) &= \sqrt{NT}(\widehat{\beta}_{NT} - \overline{\beta}_{NT}) \\
&\quad + \sqrt{NT}(\overline{B}_\infty^\beta/T + \overline{D}_\infty^\beta/N + o_P(T^{-1} \vee N^{-1})) \\
&\to_d \mathcal{N}(\kappa \overline{B}_\infty^\beta + \kappa^{-1}\overline{D}_\infty^\beta, \overline{V}_\infty).
\end{aligned}
\tag{3.3}
$$

Relative to fixed effects estimators with only individual effects, the presence of time effects introduces additional asymptotic bias through $\overline{D}_\infty^\beta$. This asymptotic result predicts that the fixed effects estimator can have significant bias relative to its dispersion. Moreover, confidence intervals constructed around the fixed effects estimator can severely undercover the true value of the parameter even in large samples. We show that these predictions provide a good approximations to the finite sample behavior of the fixed effects estimator through analytical and simulation examples in Sections 3.3 and 5.

The analytical bias correction consists of subtracting estimators of the leading terms of the bias from the fixed effect estimator of β. Let $\widehat{B}_{NT}^\beta$ and $\widehat{D}_{NT}^\beta$ be estimators of $\overline{B}_\infty^\beta$ and $\overline{D}_\infty^\beta$ as defined in (4.6). The bias corrected estimator can be formed as

$$
\widetilde{\beta}_{NT}^A = \widehat{\beta}_{NT} - \widehat{B}_{NT}^\beta/T - \widehat{D}_{NT}^\beta/N.
$$

If $N/T \to \kappa^2$, $\widehat{B}_{NT}^\beta \to_P \overline{B}_\infty^\beta$, and $\widehat{D}_{NT}^\beta \to_P \overline{D}_\infty^\beta$, then

$$
\sqrt{NT}(\widetilde{\beta}_{NT}^A - \beta^0) \to_d \mathcal{N}(0, \overline{V}_\infty).
$$

The analytical correction therefore centers the asymptotic distribution at the true value of the parameter, without increasing asymptotic variance. This asymptotic result predicts that in large samples the corrected estimator has small bias relative to dispersion, the correction does not increase dispersion, and the confidence intervals constructed around the corrected estimator have coverage probabilities close to the nominal levels. We show that these predictions provide a good approximations to the behavior of the corrections in Sections 3.3 and 5 even in small panels with $N < 60$ and $T < 15$.

We also consider a jackknife bias correction method that does not require explicit estimation of the bias. This method is based on the split panel jackknife (SPJ) of Dhaene and Jochmans (2015) applied to the time and cross-section dimension of the panel. Alternative jackknife corrections based on the leave-one-observation-out panel jackknife (PJ) of Hahn and Newey (2004) and combinations of PJ and SPJ are also possible. We do not consider corrections based on PJ because they are theoretically justified by second-order expansions of $\overline{\beta}_{NT}$ that are beyond the scope of this paper.

To describe our generalization of the SPJ, define the fixed effects estimator of β in the subpanel with cross sectional indices A and time series indices B as

$$
\begin{aligned}
\widehat{\beta}_{A,B} \in \underset{\beta \in \mathbb{R}^{\dim \beta}}{\mathrm{argmax}} \ &\max_{\alpha(A) \in \mathbb{R}^{|A|}} \\
&\times \max_{\gamma(B) \in \mathbb{R}^{|B|}} \sum_{i \in A, t \in B} \log f_Y(Y_{it} \mid X_{it}, \alpha_i, \gamma_t, \beta),
\end{aligned}
$$

where $\alpha(A) = (\alpha_i : i \in A)$ and $\gamma(B) = (\gamma_t : t \in B)$. Let $\widetilde{\beta}_{N,T/2}$ be the average of the 2 split jackknife estimators in the subpanels with $A = \{1, 2, \ldots, N\}$, and $B = \{1, 2, \ldots, T/2\}$ or $B = \{T/2 + 1, T/2 + 2, \ldots, T\}$, i.e. including all the individuals and leaving out the first and second halves of the time periods. Let $\widetilde{\beta}_{N/2,T}$ be the average of the 2 split jackknife estimators in the subpanels with $B = \{1, 2, \ldots, T\}$, and $A = \{1, 2, \ldots, N/2\}$ or $A = \{N/2 + 1, N/2 + 2, \ldots, N\}$, i.e. including all the time periods

and leaving out half of the individuals of the panel.[7] In choosing the cross sectional indexing of the panel, one might want to take into account individual clustering structures and other dependences to preserve them in the SPJ. For example, all the individuals belonging to the same cluster should be indexed such that they remain in the same subpanel after the cross sectional split. If there are no cross sectional dependences, the indexing of the individuals is unrestricted. We recommend to construct $\widetilde{\beta}_{N/2,T}$ as the average of the estimators obtained from all possible partitions of $N/2$ individuals to avoid ambiguity and arbitrariness in the choice of the division.[8] The bias corrected estimator is

$$
\widetilde{\beta}_{NT}^J = 3\widehat{\beta}_{NT} - \widetilde{\beta}_{N,T/2} - \widetilde{\beta}_{N/2,T}.
\tag{3.4}
$$

To give some intuition about how the corrections work, note that

$$
\widetilde{\beta}_{NT}^J - \beta_0 = (\widehat{\beta}_{NT} - \beta_0) - (\widetilde{\beta}_{N,T/2} - \widehat{\beta}_{NT}) - (\widetilde{\beta}_{N/2,T} - \widehat{\beta}_{NT}),
$$

where $\widetilde{\beta}_{N,T/2} - \widehat{\beta}_{NT} = \overline{B}_\infty^\beta/T + o_P(T^{-1} \vee N^{-1})$ and $\widetilde{\beta}_{N/2,T} - \widehat{\beta}_{NT} = \overline{D}_\infty^\beta/N + o_P(T^{-1} \vee N^{-1})$. Relative to $\widehat{\beta}_{NT}$, $\widetilde{\beta}_{N,T/2}$ has double the bias coming from the estimation of the individual effects because it is based on subpanels with half of the time periods, and $\widetilde{\beta}_{N/2,T}$ has double the bias coming from the estimation of the time effects because it is based on subpanels with half of the individuals. The time series split removes the bias term $\overline{B}_\infty^\beta$ and the cross sectional split removes the bias term $\overline{D}_\infty^\beta$.

3.3. Illustrative example

To illustrate how the bias corrections work in finite samples, we consider a simple model where the solution to the population program (3.1) has closed form. This model corresponds to a variation of the classical Neyman and Scott (1948) variance example that includes both individual and time effects, $Y_{it} \mid \alpha, \gamma, \beta \sim \mathcal{N}(\alpha_i + \gamma_t, \beta)$. It is well-know that in this case

$$
\widehat{\beta}_{NT} = (NT)^{-1} \sum_{i,t} \left(Y_{it} - \bar{Y}_{i.} - \bar{Y}_{.t} + \bar{Y}_{..}\right)^2,
$$

where $\bar{Y}_{i.} = T^{-1} \sum_t Y_{it}$, $\bar{Y}_{.t} = N^{-1} \sum_i Y_{it}$, and $\bar{Y}_{..} = (NT)^{-1} \sum_{i,t} Y_{it}$. Moreover, from the well-known results on the degrees of freedom adjustment of the estimated variance

$$
\begin{aligned}
\overline{\beta}_{NT} = \mathbb{E}_\phi[\widehat{\beta}_{NT}] &= \beta^0 \frac{(N-1)(T-1)}{NT} \\
&= \beta^0 \left(1 - \frac{1}{T} - \frac{1}{N} + \frac{1}{NT}\right),
\end{aligned}
$$

so that $\overline{B}_\infty^\beta = -\beta^0$ and $\overline{D}_\infty^\beta = -\beta^0$.[9]

To form the analytical bias correction we can set $\widehat{B}_{NT}^\beta = -\widehat{\beta}_{NT}$ and $\widehat{D}_{NT}^\beta = -\widehat{\beta}_{NT}$. This yields $\widetilde{\beta}_{NT}^A = \widehat{\beta}_{NT}(1 + 1/T + 1/N)$ with

$$
\overline{\beta}_{NT}^A = \mathbb{E}_\phi[\widetilde{\beta}_{NT}^A] = \beta^0 \left(1 - \frac{1}{T^2} - \frac{1}{N^2} - \frac{1}{NT} + \frac{1}{NT^2} + \frac{1}{N^2 T}\right).
$$

This correction reduces the order of the bias from $(T^{-1} \vee N^{-1})$ to $(T^{-2} \vee N^{-2})$, and introduces additional higher order terms.

---

[7] When T is odd we define $\widetilde{\beta}_{N,T/2}$ as the average of the 2 split jackknife estimators that use overlapping subpanels with $B = \{1, 2, \ldots, (T+1)/2\}$ and $B = \{(T+1)/2, (T+1)/2 + 1, \ldots, T\}$. We define $\widetilde{\beta}_{N/2,T}$ similarly when N is odd.

[8] There are $P = \binom{N}{N/2}$ different cross sectional partitions with $N/2$ individuals. When N is large, we can approximate the average over all possible partitions by the average over $S \ll P$ randomly chosen partitions to speed up computation.

[9] Okui (2013) derived the bias of fixed effects estimators of autocovariances and autocorrelations in this model.

Table 1
Biases and Standard Deviations for $Y_{it} \mid \alpha, \gamma, \beta \sim \mathcal{N}(\alpha_i + \gamma_t, \beta)$.

	$N = 10$	$N = 25$		$N = 50$		
	$T = 10$	$T = 10$	$T = 25$	$T = 10$	$T = 25$	$T = 50$
$(\overline{\beta}_{NT} - \beta^0)/\beta^0$	$-.19$	$-.14$	$-.08$	$-.12$	$-.06$	$-.04$
$(\overline{\beta}_{NT}^A - \beta^0)/\beta^0$	$-.03$	$-.02$	$.00$	$-.01$	$-.01$	$.00$
$(\overline{\beta}_{NT}^J - \beta^0)/\beta^0$	$-.01$	$.00$	$.00$	$.00$	$.00$	$.00$
$\sqrt{\overline{V}_{NT}}/\beta^0$	$.13$	$.08$	$.05$	$.06$	$.04$	$.03$
$\sqrt{\overline{V}_{NT}^A}/\beta^0$	$.14$	$.09$	$.06$	$.06$	$.04$	$.03$
$\sqrt{\overline{V}_{NT}^J}/\beta^0$	$.17$	$.10$	$.06$	$.07$	$.04$	$.03$

Notes: $\overline{V}_{NT}^J$ obtained by 50,000 simulations with $\beta^0 = 1$.

The analytical correction increases finite-sample variance because the factor $(1 + 1/T + 1/N) > 1$. We compare the biases and standard deviations of the fixed effects estimator and the corrected estimator in a numerical example below.

For the Jackknife correction, straightforward calculations give

$$\overline{\beta}_{NT}^J = \mathbb{E}_\phi[\widetilde{\beta}_{NT}^J] = 3\overline{\beta}_{NT} - \overline{\beta}_{N,T/2} - \overline{\beta}_{N/2,T} = \beta^0 \left(1 - \frac{1}{NT}\right).$$

The correction therefore reduces the order of the bias from $(T^{-1} \vee N^{-1})$ to $(TN)^{-1}$.[10]

Table 1 presents numerical results for the bias and standard deviations of the fixed effects and bias corrected estimators in finite samples. We consider panels with $N, T \in \{10, 25, 50\}$, and only report the results for $T \leq N$ since all the expressions are symmetric in N and T. All the numbers in the table are in percentage of the true parameter value, so we do not need to specify the value of β^0. We find that the analytical and jackknife corrections offer substantial improvements over the fixed effects estimator in terms of bias. The first and fourth row of the table show that the bias of the fixed effects estimator is of the same order of magnitude as the standard deviation, where $\overline{V}_{NT} = \text{Var}[\widehat{\beta}_{NT}] = 2(N-1)(T-1)(\beta^0)^2/(NT)^2$ under independence of Y_{it} over i and t conditional on the unobserved effects. The fifth row shows the increase in standard deviation due to analytical bias correction is small compared to the bias reduction, where $\overline{V}_{NT}^A = \text{Var}[\widetilde{\beta}_{NT}^A] = (1 + 1/N + 1/T)^2 \overline{V}_{NT}$. The last row shows that the jackknife yields less precise estimates than the analytical correction when $T = 10$.

Table 2 illustrates the effect of the bias on the inference based on the asymptotic distribution. It shows the coverage probabilities of 95% asymptotic confidence intervals for β^0 constructed in the usual way as

$$\text{CI}_{.95}(\widehat{\beta}) = \widehat{\beta} \pm 1.96 \widehat{V}_{NT}^{1/2} = \widehat{\beta}(1 \pm 1.96\sqrt{2/(NT)}),$$

where $\widehat{\beta} = \{\widehat{\beta}_{NT}, \widetilde{\beta}_{NT}^A, \widetilde{\beta}_{NT}^J\}$ and $\widehat{V}_{NT} = 2\widehat{\beta}^2/(NT)$ is an estimator of the asymptotic variance $\overline{V}_\infty/(NT) = 2(\beta^0)^2/(NT)$. To find the coverage probabilities, we use that $NT\widehat{\beta}_{NT}/\beta^0 \sim \chi^2_{(N-1)(T-1)}$ and $\widetilde{\beta}_{NT}^A = (1 + 1/N + 1/T)\widehat{\beta}_{NT}$. These probabilities do not depend on the value of β^0 because the limits of the intervals are proportional to $\widehat{\beta}$. For the Jackknife we compute the probabilities numerically by simulation with $\beta^0 = 1$. As a benchmark of comparison, we also consider confidence intervals constructed from the unbiased estimator $\widetilde{\beta}_{NT} = NT\widehat{\beta}_{NT}/[(N-1)(T-1)]$. Here we find that the confidence intervals based on the fixed effect estimator display

Table 2
Coverage probabilities for $Y_{it} \mid \alpha, \gamma, \beta \sim \mathcal{N}(\alpha_i + \gamma_t, \beta)$.

	$N = 10$	$N = 25$		$N = 50$		
	$T = 10$	$T = 10$	$T = 25$	$T = 10$	$T = 25$	$T = 50$
$\text{CI}_{.95}(\widehat{\beta}_{NT})$	$.56$	$.55$	$.65$	$.44$	$.63$	$.68$
$\text{CI}_{.95}(\widetilde{\beta}_{NT}^A)$	$.89$	$.92$	$.93$	$.92$	$.94$	$.94$
$\text{CI}_{.95}(\widetilde{\beta}_{NT}^J)$	$.89$	$.91$	$.93$	$.92$	$.93$	$.94$
$\text{CI}_{.95}(\widetilde{\beta}_{NT})$	$.91$	$.93$	$.94$	$.93$	$.94$	$.94$

Notes: Nominal coverage probability is .95. $\text{CI}_{.95}(\widetilde{\beta}_{NT}^J)$ obtained by 50,000 simulations with $\beta^0 = 1$.

severe undercoverage for all the sample sizes. The confidence intervals based on the corrected estimators have high coverage probabilities, which approach the nominal level as the sample size grows. Moreover, the bias corrected estimators produce confidence intervals with very similar coverage probabilities to the ones from the unbiased estimator.

4. Asymptotic theory for bias corrections

In nonlinear panel data models the population problem (3.1) generally does not have closed form solution, so we need to rely on asymptotic arguments to characterize the terms in the expansion of the bias (3.2) and to justify the validity of the corrections.

4.1. Asymptotic distribution of model parameters

We consider panel models with scalar individual and time effects that enter the likelihood function additively through $\pi_{it} = \alpha_i + \gamma_t$. In these models the dimension of the incidental parameters is $\dim \phi_{NT} = N + T$. The leading cases are single index models, where the dependence of the likelihood function on the parameters is through an index $X_{it}'\beta + \alpha_i + \gamma_t$. These models cover the probit and Poisson specifications of Examples 1 and 2. The additive structure only applies to the unobserved effects, so we can allow for scale parameters to cover the Tobit and negative binomial models. We focus on these additive models for computational and analytical tractability, because we can establish the consistency of the fixed effects estimators under a concavity assumption in the log-likelihood function with respect to all the parameters.

The parametric part of our panel models takes the form

$$\log f_Y(Y_{it} \mid X_{it}, \alpha_i, \gamma_t, \beta) =: \ell_{it}(\beta, \pi_{it}). \quad (4.1)$$

We denote the derivatives of the log-likelihood function ℓ_{it} by $\partial_\beta \ell_{it}(\beta, \pi) := \partial \ell_{it}(\beta, \pi)/\partial \beta$, $\partial_{\beta\beta'} \ell_{it}(\beta, \pi) := \partial^2 \ell_{it}(\beta, \pi)/(\partial \beta \partial \beta')$, $\partial_{\pi^q} \ell_{it}(\beta, \pi) := \partial^q \ell_{it}(\beta, \pi)/\partial \pi^q$, $q = 1, 2, 3$, *etc.* We drop the arguments β and π when the derivatives are evaluated at the true parameter values β^0 and $\pi_{it}^0 := \alpha_i^0 + \gamma_t^0$, e.g. $\partial_{\pi^q} \ell_{it} := \partial_{\pi^q} \ell_{it}(\beta^0, \pi_{it}^0)$. We also drop the dependence on NT from all the sequences of functions and parameters, e.g. we use $\mathcal{L}$ for $\mathcal{L}_{NT}$ and ϕ for ϕ_{NT}.

We make the following assumptions:

---

[10] In this example it is possible to develop higher-order jackknife corrections that completely eliminate the bias because we know the entire expansion of $\overline{\beta}_{NT}$. For example, $\mathbb{E}_\phi[4\widetilde{\beta}_{NT} - 2\widetilde{\beta}_{N,T/2} - 2\widetilde{\beta}_{N/2,T} + \widetilde{\beta}_{N/2,T/2}] = \beta^0$, where $\widetilde{\beta}_{N/2,T/2}$ is the average of the four split jackknife estimators that leave out half of the individuals and the first or the second halves of the time periods. See Dhaene and Jochmans (2015) for a discussion on higher-order bias corrections of panel fixed effects estimators.

Assumption 4.1 (*Panel Models*). Let $\nu > 0$ and $\mu > 4(8 + \nu)/\nu$. Let $\varepsilon > 0$ and let $\mathcal{B}_\varepsilon^0$ be a subset of $\mathbb{R}^{\dim \beta + 1}$ that contains an ε-neighborhood of (β^0, π_{it}^0) for all i, t, N, T.[11]

(i) Asymptotics: we consider limits of sequences where $N/T \to \kappa^2, 0 < \kappa < \infty$, as $N, T \to \infty$.

(ii) Sampling: conditional on ϕ, $\{(Y_i^T, X_i^T) : 1 \leq i \leq N\}$ is independent across i and, for each i, $\{(Y_{it}, X_{it}) : 1 \leq t \leq T\}$ is α-mixing with mixing coefficients satisfying $\sup_i a_i(m) = \mathcal{O}(m^{-\mu})$ as $m \to \infty$, where

$$a_i(m) := \sup_t \sup_{A \in \mathcal{A}_t^i, B \in \mathcal{B}_{t+m}^i} |P(A \cap B) - P(A)P(B)|,$$

and for $Z_{it} = (Y_{it}, X_{it})$, $\mathcal{A}_t^i$ is the sigma field generated by $(Z_{it}, Z_{i,t-1}, \ldots)$, and $\mathcal{B}_t^i$ is the sigma field generated by $(Z_{it}, Z_{i,t+1}, \ldots)$.

(iii) Model: for $X_i^t = \{X_{is} : s = 1, \ldots, t\}$, we assume that for all i, t, N, T,

$$Y_{it} \mid X_i^t, \phi, \beta \sim \exp[\ell_{it}(\beta, \alpha_i + \gamma_t)].$$

The realizations of the parameters and unobserved effects that generate the observed data are denoted by β^0 and ϕ^0.

(iv) Smoothness and moments: We assume that $(\beta, \pi) \mapsto \ell_{it}(\beta, \pi)$ is four times continuously differentiable over $\mathcal{B}_\varepsilon^0$ a.s. The partial derivatives of $\ell_{it}(\beta, \pi)$ with respect to the elements of (β, π) up to fourth order are bounded in absolute value uniformly over $(\beta, \pi) \in \mathcal{B}_\varepsilon^0$ by a function $M(Z_{it}) > 0$ a.s., and $\max_{i,t} \mathbb{E}_\phi[M(Z_{it})^{8+\nu}]$ is a.s. uniformly bounded over N, T.

(v) Concavity: For all N, T, $(\beta, \phi) \mapsto \mathcal{L}(\beta, \phi) = (NT)^{-1/2}\{\sum_{i,t} \ell_{it}(\beta, \alpha_i + \gamma_t) - b(v'\phi)^2/2\}$ is strictly concave over $\mathbb{R}^{\dim \beta + N + T}$ a.s. Furthermore, there exist constants $b_{\min}$ and $b_{\max}$ such that for all $(\beta, \pi) \in \mathcal{B}_\varepsilon^0$, $0 < b_{\min} \leq -\mathbb{E}_\phi\left[\partial_{\pi^2} \ell_{it}(\beta, \pi)\right] \leq b_{\max}$ a.s. uniformly over i, t, N, T.

Remark 1 (*Assumption 4.1*). Assumption 4.1(i) defines the large-T asymptotic framework and is the same as in Hahn and Kuersteiner (2011). The relative rate of N and T exactly balances the order of the bias and variance producing a non-degenerate asymptotic distribution.

Assumption 4.1(ii) does not impose identical distribution nor stationarity over the time series dimension, conditional on the unobserved effects, unlike most of the large-T panel literature, e.g., Hahn and Newey (2004) and Hahn and Kuersteiner (2011). These assumptions are violated by the presence of the time effects, because they are treated as parameters. The mixing condition is used to bound covariances and moments in the application of laws of large numbers and central limit theorems—it could replaced by other conditions that guarantee the applicability of these results.

Assumption 4.1(iii) is the parametric part of the panel model. We rely on this assumption to guarantee that $\partial_\beta \ell_{it}$ and $\partial_\pi \ell_{it}$ have martingale difference properties. Moreover, we use certain Bartlett identities implied by this assumption to simplify some expressions, but those simplifications are not crucial for our results. We provide expressions for the asymptotic bias and variance that do not apply these simplifications in Remark 3.

Assumption 4.1(iv) imposes smoothness and moment conditions in the log-likelihood function and its derivatives. These conditions guarantee that the higher-order stochastic expansions of the fixed effect estimator that we use to characterize the asymptotic bias are well-defined, and that the remainder terms of these expansions are bounded.

The most commonly used nonlinear models in applied economics such as logit, probit, ordered probit, Poisson, and Tobit models have smooth log-likelihoods functions that satisfy the concavity condition of Assumption 4.1(v), provided that all the elements of X_{it} have cross sectional and time series variation. Assumption 4.1(v) guarantees that β^0 and ϕ^0 are the unique solution to the population problem (2.1), that is all the parameters are point identified.

To describe the asymptotic distribution of the fixed effects estimator $\widehat{\beta}$, it is convenient to introduce some additional notation. Let $\overline{\mathcal{H}}$ be the $(N + T) \times (N + T)$ expected Hessian matrix of the log-likelihood with respect to the nuisance parameters evaluated at the true parameters, i.e.

$$\overline{\mathcal{H}} = \mathbb{E}_\phi[-\partial_{\phi\phi'}\mathcal{L}] = \begin{pmatrix} \overline{\mathcal{H}}_{(\alpha\alpha)}^* & \overline{\mathcal{H}}_{(\alpha\gamma)}^* \\ [\overline{\mathcal{H}}_{(\alpha\gamma)}^*]' & \overline{\mathcal{H}}_{(\gamma\gamma)}^* \end{pmatrix} + \frac{b}{\sqrt{NT}} vv', \quad (4.2)$$

where $\overline{\mathcal{H}}_{(\alpha\alpha)}^* = \mathrm{diag}(\sum_t \mathbb{E}_\phi[-\partial_{\pi^2}\ell_{it}])/\sqrt{NT}$, $\overline{\mathcal{H}}_{(\alpha\gamma)it}^* = \mathbb{E}_\phi[-\partial_{\pi^2}\ell_{it}]/\sqrt{NT}$, and $\overline{\mathcal{H}}_{(\gamma\gamma)}^* = \mathrm{diag}(\sum_i \mathbb{E}_\phi[-\partial_{\pi^2}\ell_{it}])/\sqrt{NT}$. Furthermore, let $\overline{\mathcal{H}}_{(\alpha\alpha)}^{-1}, \overline{\mathcal{H}}_{(\alpha\gamma)}^{-1}, \overline{\mathcal{H}}_{(\gamma\alpha)}^{-1}$, and $\overline{\mathcal{H}}_{(\gamma\gamma)}^{-1}$ denote the $N \times N$, $N \times T, T \times N$ and $T \times T$ blocks of the inverse $\overline{\mathcal{H}}^{-1}$ of $\overline{\mathcal{H}}$. We define the $\dim \beta$-vector Ξ_{it} and the operator $D_{\beta\pi^q}$ as

$$\Xi_{it} := -\frac{1}{\sqrt{NT}} \sum_{j=1}^N \sum_{\tau=1}^T \left(\overline{\mathcal{H}}_{(\alpha\alpha)ij}^{-1} + \overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1} \right.$$
$$\left. + \overline{\mathcal{H}}_{(\alpha\gamma)i\tau}^{-1} + \overline{\mathcal{H}}_{(\gamma\gamma)t\tau}^{-1} \right) \mathbb{E}_\phi \left(\partial_{\beta\pi} \ell_{j\tau} \right),$$

$$D_{\beta\pi^q} \ell_{it} := \partial_{\beta\pi^q} \ell_{it} - \partial_{\pi^{q+1}} \ell_{it} \Xi_{it}, \quad (4.3)$$

with $q = 0, 1, 2$. The kth component of Ξ_{it} corresponds to the population least squares projection of $\mathbb{E}_\phi(\partial_{\beta_k\pi}\ell_{it})/\mathbb{E}_\phi(\partial_{\pi^2}\ell_{it})$ on the space spanned by the incidental parameters under a metric given by $\mathbb{E}_\phi(-\partial_{\pi^2}\ell_{it})$, i.e.

$$\Xi_{it,k} = \alpha_{i,k}^* + \gamma_{t,k}^*,$$
$$(\alpha_k^*, \gamma_k^*) = \underset{\alpha_{i,k}, \gamma_{t,k}}{\mathrm{argmin}} \sum_{i,t} \mathbb{E}_\phi(-\partial_{\pi^2}\ell_{it})$$
$$\times \left(\frac{\mathbb{E}_\phi(\partial_{\beta_k\pi}\ell_{it})}{\mathbb{E}_\phi(\partial_{\pi^2}\ell_{it})} - \alpha_{i,k} - \gamma_{t,k} \right)^2.$$

The operator $D_{\beta\pi^q}$ partials out individual and time effects in non-linear models. It corresponds to individual and time differencing when the model is linear. To see this, consider the normal linear model $Y_{it} \mid X_i^t, \alpha_i, \gamma_t \sim \mathcal{N}(X_{it}'\beta + \alpha_i + \gamma_t, 1)$. Then, $\Xi_{it} = T^{-1}\sum_{t=1}^T \mathbb{E}_\phi[X_{it}] + N^{-1}\sum_{i=1}^N \mathbb{E}_\phi[X_{it}] - (NT)^{-1}\sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_\phi[X_{it}]$, $D_\beta \ell_{it} = -\tilde{X}_{it}\varepsilon_{it}, D_{\beta\pi} \ell_{it} = -\tilde{X}_{it}$, and $D_{\beta\pi^2} \ell_{it} = 0$, where $\varepsilon_{it} = Y_{it} - X_{it}'\beta - \alpha_i - \gamma_t$ and $\tilde{X}_{it} = X_{it} - \Xi_{it}$ is the vector of individual and time demeaned explanatory variables.

The following theorem establishes the asymptotic distribution of the fixed effects estimator $\widehat{\beta}$.

Theorem 4.1 (*Asymptotic Distribution of* $\widehat{\beta}$). *Suppose that Assumption 4.1 holds, that the following limits exist*

$$\overline{B}_\infty$$

$$= \underset{N,T \to \infty}{\mathrm{plim}} \left[-\frac{1}{N} \sum_{i=1}^N \frac{\sum_{t=1}^T \sum_{\tau=t}^T \mathbb{E}_\phi \left(\partial_\pi \ell_{it} D_{\beta\pi} \ell_{i\tau} \right) + \frac{1}{2} \sum_{t=1}^T \mathbb{E}_\phi(D_{\beta\pi^2} \ell_{it})}{\sum_{t=1}^T \mathbb{E}_\phi \left(\partial_{\pi^2} \ell_{it} \right)} \right],$$

---

[11] For example, $\mathcal{B}_\varepsilon^0$ can be chosen to be the Cartesian product of the ε-ball around β^0 and the interval $[\pi_{\min}, \pi_{\max}]$, with $\pi_{\min} \leq \pi_{it} - \varepsilon$ and $\pi_{\max} \geq \pi_{it} + \varepsilon$ for all i, t, N, T. We can have $\pi_{\min} = -\infty$ and $\pi_{\max} = \infty$, as long as this is compatible with Assumption 4.1(iv) and (v).

$$\overline{D}_\infty = \plim_{N,T\to\infty} \left[-\frac{1}{T}\sum_{t=1}^{T} \frac{\sum_{i=1}^{N} \mathbb{E}_\phi\left(\partial_\pi \ell_{it} D_{\beta\pi}\ell_{it} + \frac{1}{2}D_{\beta\pi^2}\ell_{it}\right)}{\sum_{i=1}^{N}\mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)} \right],$$

$$\overline{W}_\infty = \plim_{N,T\to\infty} \left[-\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}_\phi\left(\partial_{\beta\beta'}\ell_{it} - \partial_{\pi^2}\ell_{it}\,\Xi_{it}\,\Xi_{it}'\right) \right],$$

and that $\overline{W}_\infty > 0$. *Then,*

$$\sqrt{NT}\left(\widehat{\beta} - \beta^0\right) \to_d \overline{W}_\infty^{-1}\mathcal{N}(\kappa\overline{B}_\infty + \kappa^{-1}\overline{D}_\infty, \overline{W}_\infty),$$

so that $\overline{B}_\infty^\beta = \overline{W}_\infty^{-1}\overline{B}_\infty, \overline{D}_\infty^\beta = \overline{W}_\infty^{-1}\overline{D}_\infty$, *and* $\overline{V}_\infty = \overline{W}_\infty^{-1}$ *in* (3.2) *and* (3.3).

Remark 2. The complete proof of Theorem 4.1 is provided in the Appendix. Here we point out why the argument for the consistency proof in models with only individual effects does not apply to our setting, give a heuristic derivation of the asymptotic distribution, and highlight where some of the assumptions are used in the proof.

(i) The consistency proof for models with only individual effects relies on partitioning the log-likelihood in the sum of individual log-likelihoods that depend on a fixed number of parameters, the model parameter β and the corresponding individual effect α_i. The maximizers of the individual log-likelihood are then consistent estimators of all the parameters as T becomes large by standard arguments. This approach does not work in models with individual and time effects because there is no partition of the data that is only affected by a fixed number of parameters, and whose size grows with the sample size.

(ii) In the following we give a heuristic discussion of the asymptotic distribution result for $\widehat{\beta}$. A first-order Taylor series expansion to approximate the first order conditions of (2.8) around β^0 gives

$$0 = \partial_\beta\mathcal{L}(\widehat{\beta}, \widehat{\phi}(\widehat{\beta})) \approx \partial_\beta\mathcal{L}(\beta^0, \widehat{\phi}^0)$$
$$- \overline{W}_\infty\sqrt{NT}(\widehat{\beta} - \beta^0), \qquad (4.4)$$

where $\widehat{\phi}^0 = \widehat{\phi}(\beta^0)$. A second-order Taylor series expansion to approximate $\partial_\beta\mathcal{L}(\beta^0, \widehat{\phi}^0)$ around ϕ^0 yields

$$\partial_\beta\mathcal{L}(\beta^0, \widehat{\phi}^0) \approx \partial_\beta\mathcal{L}(\beta^0, \phi^0) + \partial_{\beta\phi'}\mathcal{L}(\beta^0, \phi^0)[\widehat{\phi}^0 - \phi^0]$$
$$+ \sum_{g=1}^{\dim\phi} \partial_{\beta\phi'\phi_g}\mathcal{L}(\beta^0, \phi^0)[\widehat{\phi}^0 - \phi^0][\widehat{\phi}_g^0 - \phi_g^0]/2,$$

where the first term has zero mean and determines the asymptotic variance, and the second and third term determine the asymptotic bias. Thus, by the central limit theorem and the information equality,

$$\partial_\beta\mathcal{L}(\beta^0, \phi^0) \to_d \mathcal{N}(0, \overline{W}_\infty).$$

The second and third terms satisfy

$$\partial_{\beta\phi'}\mathcal{L}(\beta^0, \phi^0)[\widehat{\phi}^0 - \phi^0]$$
$$+ \sum_{g=1}^{\dim\phi} \partial_{\beta\phi'\phi_g}\mathcal{L}(\beta^0, \phi^0)[\widehat{\phi}^0 - \phi^0][\widehat{\phi}_g^0 - \phi_g^0]/2$$
$$\approx \sqrt{NT}(\overline{B}_\infty/T + \overline{D}_\infty/N),$$

where $\overline{B}_\infty$ and $\overline{D}_\infty$ are characterized from a second-order Taylor series expansion to approximate $\widehat{\phi}^0$ around ϕ^0. We refer to the Appendix for the details of this derivation. There we show that $\overline{B}_\infty$ and $\overline{D}_\infty$ originate from the elements of $\widehat{\phi}^0$ corresponding to the individual effects and time effects,

respectively. Plugging those results into (4.4), and solving for $\sqrt{NT}(\widehat{\beta} - \beta^0)$ yields

$$\sqrt{NT}(\widehat{\beta} - \beta^0) \approx \overline{W}_\infty^{-1}[\partial_\beta\mathcal{L}(\beta^0, \phi^0) + \overline{B}_\infty\sqrt{N/T}$$
$$+ \overline{D}_\infty\sqrt{T/N}] \to_d \overline{W}_\infty^{-1}\mathcal{N}(\kappa\overline{B}_\infty + \kappa^{-1}\overline{D}_\infty, \overline{W}_\infty).$$

This derivation shows that the source of the bias is that the score $\partial_\beta\mathcal{L}(\beta, \widehat{\phi}(\beta))$ is not centered at zero when $\beta = \beta^0$. This problem arises from the substitution of the incidental parameter ϕ by the sample analog $\widehat{\phi}^0$ that has a rate of convergence slower than $\sqrt{NT}$. Thus, $\overline{B}_\infty$ originates from the estimators of the individual effects in ϕ, which have rate of convergence $\sqrt{T}$; whereas $\overline{D}_\infty$ originates from the estimators of the time effects in ϕ, which have convergence rate $\sqrt{N}$.

(iii) The two key assumptions in the derivation of the asymptotic distribution are the additive separability of α_i and γ_t in Assumption 4.1(iii) and the concavity in Assumption 4.1(v). We resort to concavity to prove consistency of $\widehat{\beta}$ and to bound the remainder terms in all the expansions. Additive separability is convenient to characterize the order of the inverse average Hessian, $\overline{\mathcal{H}}$, defined in (4.2). This inverse Hessian features prominently in the second-order Taylor series expansion of $\widehat{\phi}^0$ around ϕ^0 used to characterize $\overline{B}_\infty$ and $\overline{D}_\infty$.

It is instructive to evaluate the expressions of the bias in our running examples.

Example 1 (*Binary Response Model*). In this case

$$\ell_{it}(\beta, \pi) = Y_{it}\log F(X_{it}'\beta + \pi) + (1 - Y_{it})\log[1 - F(X_{it}'\beta + \pi)],$$

so that $\partial_\pi\ell_{it} = H_{it}(Y_{it} - F_{it})$, $\partial_\beta\ell_{it} = \partial_\pi\ell_{it}X_{it}$, $\partial_{\pi^2}\ell_{it} = -H_{it}\partial F_{it} + \partial H_{it}(Y_{it} - F_{it})$, $\partial_{\beta\beta'}\ell_{it} = \partial_{\pi^2}\ell_{it}X_{it}X_{it}'$, $\partial_{\beta\pi}\ell_{it} = \partial_{\pi^2}\ell_{it}X_{it}$, $\partial_{\pi^3}\ell_{it} = -H_{it}\partial^2 F_{it} - 2\partial H_{it}\partial F_{it} + \partial^2 H_{it}(Y_{it} - F_{it})$, and $\partial_{\beta\pi^2}\ell_{it} = \partial_{\pi^3}\ell_{it}X_{it}$, where $H_{it} = \partial F_{it}/[F_{it}(1 - F_{it})]$, *and* $\partial^j G_{it} := \partial^j G(Z)|_{Z=X_{it}'\beta^0 + \pi_{it}^0}$ for any function G and $j = 0, 1, 2$. Substituting these values in the expressions of the bias of Theorem 4.1 yields

$$\overline{B}_\infty = \plim_{N,T\to\infty}\left[-\frac{1}{2N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T}\mathbb{E}_\phi[H_{it}\partial^2 F_{it}\tilde{X}_{it}]}{\sum_{t=1}^{T}\mathbb{E}_\phi(\omega_{it})} \right.$$
$$\left. -\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T}\sum_{\tau=t+1}^{T}\mathbb{E}_\phi\left[H_{it}(Y_{it} - F_{it})\omega_{i\tau}\tilde{X}_{i\tau}\right]}{\sum_{t=1}^{T}\mathbb{E}_\phi(\omega_{it})} \right],$$

$$\overline{D}_\infty = \plim_{N,T\to\infty}\left[-\frac{1}{2T}\sum_{t=1}^{T}\frac{\sum_{i=1}^{N}\mathbb{E}_\phi[H_{it}\partial^2 F_{it}\tilde{X}_{it}]}{\sum_{i=1}^{N}\mathbb{E}_\phi(\omega_{it})} \right],$$

$$\overline{W}_\infty = \plim_{N,T\to\infty}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}_\phi[\omega_{it}\tilde{X}_{it}\tilde{X}_{it}'] \right],$$

where $\omega_{it} = H_{it}\partial F_{it}$ and $\tilde{X}_{it}$ is the residual of the population projection of X_{it} on the space spanned by the incidental parameters under a metric weighted by $\mathbb{E}_\phi(\omega_{it})$. For the probit model where all the components of X_{it} are strictly exogenous,

$$\overline{B}_\infty = \plim_{N,T\to\infty}\left[\frac{1}{2N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T}\mathbb{E}_\phi[\omega_{it}\tilde{X}_{it}\tilde{X}_{it}']}{\sum_{t=1}^{T}\mathbb{E}_\phi(\omega_{it})} \right]\beta^0,$$

$$\overline{D}_\infty = \plim_{N,T\to\infty} \left[\frac{1}{2T} \sum_{t=1}^T \frac{\sum_{i=1}^N \mathbb{E}_\phi[\omega_{it}\tilde{X}_{it}\tilde{X}_{it}']}{\sum_{i=1}^N \mathbb{E}_\phi(\omega_{it})} \right] \beta^0.$$

The asymptotic bias is therefore a positive definite matrix weighted average of the true parameter value as in the case of the probit model with only individual effects (Fernández-Val, 2009).

Example 2 (*Poisson Model*). In this case

$$\ell_{it}(\beta, \pi) = (X_{it}'\beta + \pi)Y_{it} - \exp(X_{it}'\beta + \pi) - \log Y_{it}!,$$

so that $\partial_\pi \ell_{it} = Y_{it} - \omega_{it}$, $\partial_\beta \ell_{it} = \partial_\pi \ell_{it} X_{it}$, $\partial_{\pi^2} \ell_{it} = \partial_{\pi^3}\ell_{it} = -\omega_{it}$, $\partial_{\beta\beta'}\ell_{it} = \partial_{\pi^2}\ell_{it}X_{it}X_{it}'$, and $\partial_{\beta\pi}\ell_{it} = \partial_{\beta\pi^2}\ell_{it} = \partial_{\pi^3}\ell_{it}X_{it}$, where $\omega_{it} = \exp(X_{it}'\beta^0 + \pi_{it}^0)$. Substituting these values in the expressions of the bias of Theorem 4.1 yields

$$\overline{B}_\infty = \plim_{N,T\to\infty} \left[-\frac{1}{N} \sum_{i=1}^N \frac{\sum_{t=1}^T \sum_{\tau=t+1}^T \mathbb{E}_\phi\left[(Y_{it} - \omega_{it})\omega_{i\tau}\tilde{X}_{i\tau}\right]}{\sum_{t=1}^T \mathbb{E}_\phi(\omega_{it})} \right],$$

$$\overline{W}_\infty = \plim_{N,T\to\infty} \left[\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_\phi[\omega_{it}\tilde{X}_{it}\tilde{X}_{it}'] \right],$$

and $\overline{D}_\infty = 0$, where $\tilde{X}_{it}$ is the residual of the population projection of X_{it} on the space spanned by the incidental parameters under a metric weighted by $\mathbb{E}_\phi(\omega_{it})$. If in addition all the components of X_{it} are strictly exogenous, then we get the no asymptotic bias result $\overline{B}_\infty = \overline{D}_\infty = 0$.

Remark 3 (*Bias and Variance Expressions for Conditional Moment Models*). In the derivation of the asymptotic distribution, we apply Bartlett identities implied by Assumption 4.1(iii) to simplify the expressions. The following expressions of the asymptotic bias and variance do not make use of these identities and therefore remain valid in conditional moment models that do not specify the entire conditional distribution of Y_{it}:

$$\overline{B}_\infty = \plim_{N,T\to\infty} \left[-\frac{1}{N}\sum_{i=1}^N \frac{\sum_{t=1}^T \sum_{\tau=t}^T \mathbb{E}_\phi\left(\partial_\pi \ell_{it}D_{\beta\pi}\ell_{i\tau}\right)}{\sum_{t=1}^T \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)} \right]$$
$$+ \frac{1}{2}\plim_{N,T\to\infty}\left[\frac{1}{N}\sum_{i=1}^N \frac{\sum_{t=1}^T \mathbb{E}_\phi[(\partial_\pi \ell_{it})^2]\sum_{t=1}^T \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it})}{\left[\sum_{t=1}^T \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)\right]^2} \right],$$

$$\overline{D}_\infty = \plim_{N,T\to\infty}\left[-\frac{1}{T}\sum_{t=1}^T \frac{\sum_{i=1}^N \mathbb{E}_\phi\left[\partial_\pi \ell_{it}D_{\beta\pi}\ell_{it}\right]}{\sum_{i=1}^N \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)} \right]$$
$$+ \frac{1}{2}\plim_{N,T\to\infty}\left[\frac{1}{T}\sum_{t=1}^T \frac{\sum_{i=1}^N \mathbb{E}_\phi[(\partial_\pi \ell_{it})^2]\sum_{i=1}^N \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it})}{\left[\sum_{i=1}^N \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)\right]^2} \right],$$

$$\overline{V}_\infty = \overline{W}_\infty^{-1}\overline{\Omega}_\infty\overline{W}_\infty^{-1},$$

$$\overline{\Omega}_\infty = \plim_{N,T\to\infty}\left[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \sum_{\tau=1}^T \mathbb{E}_\phi\left[D_\beta \ell_{it}(D_\beta \ell_{i\tau})'\right] \right],$$

and $\overline{W}_\infty$ is the same as in Theorem 4.1.

For example, consider the Poisson fixed effects estimator in the conditional mean model $\mathbb{E}[Y_{it} \mid X_i^t, \phi, \beta] = \omega_{it} = \exp(X_{it}'\beta + \alpha_i + \gamma_t)$. Applying the previous expressions to $\ell_{it}(\beta, \pi) = (X_{it}'\beta + \pi)Y_{it} - \exp(X_{it}'\beta + \pi) - \log Y_{it}!$ yields the same expressions for $\overline{B}_\infty, \overline{D}_\infty, \overline{W}_\infty$ as in Example 2, and

$$\overline{\Omega}_\infty = \plim_{N,T\to\infty}\left[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_\phi\left[(Y_{it} - \omega_{it})^2\tilde{X}_{it}\tilde{X}_{it}'\right] \right],$$

where $\tilde{X}_{it}$ is defined as in Example 2. If all the components of X_{it} are strictly exogenous, then we get again the no asymptotic bias result $\overline{B}_\infty = \overline{D}_\infty = 0$.

4.2. Asymptotic distribution of APEs

In nonlinear models we are often interested in APEs, in addition to model parameters. These effects are averages of the data, parameters and unobserved effects; see expression (2.2). For the panel models of Assumption 4.1 we specify the partial effects as $\Delta(X_{it}, \beta, \alpha_i, \gamma_t) = \Delta_{it}(\beta, \pi_{it})$. The restriction that the partial effects depend on α_i and γ_t through π_{it} is natural in our panel models since

$$\mathbb{E}[Y_{it} \mid X_i^t, \alpha_i, \gamma_t, \beta] = \int y \exp[\ell_{it}(\beta, \pi_{it})]dy,$$

and the partial effects are usually defined as differences or derivatives of this conditional expectation with respect to the components of X_{it}. For example, the partial effects for the binary response and Poisson models described in Section 2 satisfy this restriction.

The distribution of the unobserved individual and time effects is not ancillary for the APEs, unlike for model parameters. We therefore need to make assumptions on this distribution to define and interpret the APEs, and to derive the asymptotic distribution of their estimators. We control the heterogeneity of the partial effects assuming that the individual effects and explanatory variables are identically distributed cross sectionally and/or stationary over time. If $(X_{it}, \alpha_i, \gamma_t)$ is identically distributed over i but can be heterogeneously distributed over t, then $\mathbb{E}[\Delta_{it}] = \delta_t^0$, and $\delta_{NT}^0 = T^{-1}\sum_{t=1}^T \delta_t^0$ changes only with T. If $(X_{it}, \alpha_i, \gamma_t)$ is stationary over t but can be heterogeneously distributed over i, then $\mathbb{E}[\Delta_{it}] = \delta_i^0$, and $\delta_{NT}^0 = N^{-1}\sum_{i=1}^N \delta_i^0$ changes only with N. Finally, if $(X_{it}, \alpha_i, \gamma_t)$ is identically distributed over i and stationary over t, then $\mathbb{E}[\Delta_{it}] = \delta_{NT}^0$, and $\delta_{NT}^0 = \delta^0$ does not change with N and T. We also impose smoothness and moment conditions on the function Δ that defines the partial effects. We use these conditions to derive higher-order stochastic expansions for the fixed effect estimator of the APEs and to bound the remainder terms in these expansions. Let $\{\alpha_i\}_N := \{\alpha_i : 1 \le i \le N\}$, $\{\gamma_t\}_T := \{\gamma_t : 1 \le t \le T\}$, and $\{X_{it}, \alpha_i, \gamma_t\}_{NT} := \{(X_{it}, \alpha_i, \gamma_t) : 1 \le i \le N, 1 \le t \le T\}$.

Assumption 4.2 (*Partial Effects*). Let $\nu > 0$, $\epsilon > 0$, and $\mathcal{B}_\epsilon^0$ all be as in Assumption 4.1.

(i) Sampling: for all N, T, $\{X_{it}, \alpha_i, \gamma_t\}_{NT}$ is identically distributed across i and/or stationary across t.[12]

(ii) Model: for all i, t, N, T, the partial effects depend on α_i and γ_t through $\alpha_i + \gamma_t$:

$$\Delta(X_{it}, \beta, \alpha_i, \gamma_t) = \Delta_{it}(\beta, \alpha_i + \gamma_t).$$

The realizations of the partial effects are denoted by $\Delta_{it} := \Delta_{it}(\beta^0, \alpha_i^0 + \gamma_t^0)$.

---

[12] In the working paper version, Fernández-Val and Weidner (2015a), we also consider inference conditional on the unobserved effects by assuming that $\{\alpha_i\}_N$ and $\{\gamma_t\}_T$ are deterministic sequences.

(iii) Smoothness and moments: The function $(\beta, \pi) \mapsto \Delta_{it}(\beta, \pi)$ is four times continuously differentiable over $\mathcal{B}^0_\varepsilon$ a.s. The partial derivatives of $\Delta_{it}(\beta, \pi)$ with respect to the elements of (β, π) up to fourth order are bounded in absolute value uniformly over $(\beta, \pi) \in \mathcal{B}^0_\varepsilon$ by a function $M(Z_{it}) > 0$ a.s., and $\max_{i,t} \mathbb{E}_\phi[M(Z_{it})^{8+\nu}]$ is a.s. uniformly bounded over N, T.

(iv) Non-degeneracy and moments: $0 < \min_{i,t}[\mathbb{E}(\Delta_{it}^2) - \mathbb{E}(\Delta_{it})^2] \leq \max_{i,t}[\mathbb{E}(\Delta_{it}^2) - \mathbb{E}(\Delta_{it})^2] < \infty$, uniformly over N, T.

Analogous to Ξ_{it} in Eq. (4.3) we define

$$\Psi_{it} = -\frac{1}{\sqrt{NT}} \sum_{j=1}^{N} \sum_{\tau=1}^{T} \left(\overline{\mathcal{H}}^{-1}_{(\alpha\alpha)ij} + \overline{\mathcal{H}}^{-1}_{(\gamma\alpha)tj} \right.$$
$$\left. + \overline{\mathcal{H}}^{-1}_{(\alpha\gamma)i\tau} + \overline{\mathcal{H}}^{-1}_{(\gamma\gamma)t\tau} \right) \partial_\pi \Delta_{j\tau}, \quad (4.5)$$

which is the population projection of $\partial_\pi \Delta_{it} / \mathbb{E}_\phi[\partial_{\pi^2} \ell_{it}]$ on the space spanned by the incidental parameters under the metric given by $\mathbb{E}_\phi[-\partial_{\pi^2} \ell_{it}]$. We use analogous notation to the previous section for the derivatives with respect to β and higher order derivatives with respect to π.

Let δ^0_{NT} and $\widehat{\delta}$ be the APE and its fixed effects estimator, defined as in Eqs. (2.2) and (2.9) with $\Delta(X_{it}, \beta, \alpha_i, \gamma_t) = \Delta_{it}(\beta, \alpha_i + \gamma_t)$.[13] The following theorem establishes the asymptotic distribution of $\widehat{\delta}$.

Theorem 4.2 (*Asymptotic Distribution of $\widehat{\delta}$*)**.** *Suppose that the assumptions of Theorem 4.1 and Assumption 4.2 hold, and that the following limits exist:*

$$\overline{(D_\beta \Delta)}_\infty = \operatorname*{plim}_{N,T\to\infty} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}_\phi (\partial_\beta \Delta_{it} - \Xi_{it} \partial_\pi \Delta_{it}) \right],$$

$$\overline{B}^\delta_\infty = \overline{(D_\beta \Delta)}'_\infty \overline{W}^{-1}_\infty \overline{B}_\infty$$
$$+ \operatorname*{plim}_{N,T\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{t=1}^{T} \sum_{\tau=t}^{T} \mathbb{E}_\phi \left(\partial_\pi \ell_{it} \partial_{\pi^2} \ell_{i\tau} \Psi_{i\tau} \right)}{\sum_{t=1}^{T} \mathbb{E}_\phi \left(\partial_{\pi^2} \ell_{it} \right)} \right]$$
$$- \operatorname*{plim}_{N,T\to\infty} \left[\frac{1}{2N} \sum_{i=1}^{N} \frac{\sum_{t=1}^{T} \left[\mathbb{E}_\phi(\partial_{\pi^2} \Delta_{it}) - \mathbb{E}_\phi(\partial_{\pi^3} \ell_{it}) \mathbb{E}_\phi(\Psi_{it}) \right]}{\sum_{t=1}^{T} \mathbb{E}_\phi \left(\partial_{\pi^2} \ell_{it} \right)} \right],$$

$$\overline{D}^\delta_\infty = \overline{(D_\beta \Delta)}'_\infty \overline{W}^{-1}_\infty \overline{D}_\infty$$
$$+ \operatorname*{plim}_{N,T\to\infty} \left[\frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i=1}^{N} \mathbb{E}_\phi \left(\partial_\pi \ell_{it} \partial_{\pi^2} \ell_{it} \Psi_{it} \right)}{\sum_{i=1}^{N} \mathbb{E}_\phi \left(\partial_{\pi^2} \ell_{it} \right)} \right]$$
$$- \operatorname*{plim}_{N,T\to\infty} \left[\frac{1}{2T} \sum_{t=1}^{T} \frac{\sum_{i=1}^{N} \left[\mathbb{E}_\phi(\partial_{\pi^2} \Delta_{it}) - \mathbb{E}_\phi(\partial_{\pi^3} \ell_{it}) \mathbb{E}_\phi(\Psi_{it}) \right]}{\sum_{i=1}^{N} \mathbb{E}_\phi \left(\partial_{\pi^2} \ell_{it} \right)} \right],$$

$$\overline{V}^\delta_\infty = \operatorname*{plim}_{N,T\to\infty} \left\{ \frac{r_{NT}^2}{N^2 T^2} \mathbb{E} \left[\left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{\Delta}_{it} \right) \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \widetilde{\Delta}_{it} \right)' \right. \right.$$
$$\left. \left. + \sum_{i=1}^{N} \sum_{t=1}^{T} \Gamma_{it} \Gamma'_{it} \right] \right\},$$

for some deterministic sequence $r_{NT} \to \infty$ *such that* $r_{NT} = \mathcal{O}(\sqrt{NT})$ *and* $\overline{V}^\delta_\infty > 0$, *where* $\widetilde{\Delta}_{it} = \Delta_{it} - \mathbb{E}(\Delta_{it})$ *and* $\Gamma_{it} = \overline{(D_\beta \Delta)}'_\infty \overline{W}^{-1}_\infty D_\beta \ell_{it} - \mathbb{E}_\phi(\Psi_{it}) \partial_\pi \ell_{it}$. *Then,*

$$r_{NT}(\widehat{\delta} - \delta^0_{NT} - T^{-1} \overline{B}^\delta_\infty - N^{-1} \overline{D}^\delta_\infty) \to_d \mathcal{N}(0, \overline{V}^\delta_\infty).$$

Remark 4 (*Convergence Rate, Bias and Variance*)**.** To understand the asymptotic distribution of $\widehat{\delta}$ is useful to decompose

$$\widehat{\delta} - \delta^0_{NT} = [\widehat{\delta} - \delta] + [\delta - \delta^0_{NT}],$$

where $\delta := (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \Delta_{it}$. In this decomposition the first term captures variation due to parameter estimation, whereas the second term captures variation due to estimation of a population mean by a sample mean. Under Assumption 4.2(iv) the convergence rate r_{NT} is determined by the convergence rate of $\delta - \delta^0_{NT}$, which depends on the sampling properties of the unobserved effects. For example, if $\{\alpha_i\}_N$ and $\{\gamma_t\}_T$ are independent sequences, and α_i and γ_t are independent for all i, t, then $r_{NT} = \sqrt{NT/(N+T-1)}$, and

$$\overline{V}^\delta_\infty = \operatorname*{plim}_{N,T\to\infty} \left\{ \frac{r_{NT}^2}{N^2 T^2} \sum_{i=1}^{N} \left[\sum_{t,\tau=1}^{T} \mathbb{E}(\widetilde{\Delta}_{it} \widetilde{\Delta}'_{i\tau}) + \sum_{j\neq i} \sum_{t=1}^{T} \mathbb{E}(\widetilde{\Delta}_{it} \widetilde{\Delta}'_{jt}) \right. \right.$$
$$\left. \left. + \sum_{t=1}^{T} \mathbb{E}(\Gamma_{it} \Gamma'_{it}) \right] \right\}.$$

In the expression of $\overline{V}^\delta_\infty$, the first two terms come from $\delta - \delta^0_{NT}$, whereas the last term comes from $\widehat{\delta} - \delta$. The first two terms of $\overline{V}^\delta_\infty$ are of order $NT(T + N - 1)r_{NT}^2/(NT)^2 = \mathcal{O}(1)$ by construction, the last term of $\overline{V}^\delta_\infty$ is of order $NTr_{NT}^2/(NT)^2 = \mathcal{O}(T^{-1} + N^{-1})$, and the asymptotic bias $r_{NT}(T^{-1}\overline{B}^\delta_\infty + N^{-1}\overline{D}^\delta_\infty)$ is of order $r_{NT}(T^{-1}+N^{-1}) = \mathcal{O}(T^{-1/2} + N^{-1/2})$. Thus, the bias and variance coming from parameter estimation are asymptotically negligible relative to the variances coming from the estimation of a population mean by a sample mean. In numerical examples, however, we find that correcting the mean and variance for parameter estimation improves the finite-sample estimation and inference properties of the APE estimators.

Remark 5 (*Average Effects from Bias Corrected Estimators*)**.** The first term in the expressions of the biases $\overline{B}^\delta_\infty$ and $\overline{D}^\delta_\infty$ comes from the bias of the estimator of β. It drops out when the APEs are constructed from asymptotically unbiased or bias corrected estimators of the parameter β, i.e.

$$\widetilde{\delta} = \Delta(\widetilde{\beta}, \widehat{\phi}(\widetilde{\beta})),$$

where $\widetilde{\beta}$ is such that $\sqrt{NT}(\widetilde{\beta} - \beta^0) \to_d N(0, \overline{W}^{-1}_\infty)$. The asymptotic variance of $\widetilde{\delta}$ is the same as in Theorem 4.2.

In the following examples we assume that the APEs are constructed from asymptotically unbiased estimators of the model parameters.

Example 1 (*Binary Response Model*)**.** Consider the partial effects defined in (2.3) and (2.4) with

$$\Delta_{it}(\beta, \pi) = F(\beta_k + X'_{it,-k}\beta_{-k} + \pi) - F(X'_{it,-k}\beta_{-k} + \pi) \quad \text{and}$$
$$\Delta_{it}(\beta, \pi) = \beta_k \partial F(X'_{it}\beta + \pi).$$

---

[13] We keep the dependence of δ^0_{NT} on NT to distinguish δ^0_{NT} from $\delta^0 = \lim_{N,T\to\infty} \delta^0_{NT}$.

Using the notation previously introduced for this example, the components of the asymptotic bias of $\widehat{\delta}$ are

$$\overline{B}^\delta_\infty = \underset{N,T\to\infty}{\text{plim}} \left[\frac{1}{N}\sum_{i=1}^N \frac{\sum_{t=1}^T \sum_{\tau=t+1}^T \mathbb{E}_\phi\left(H_{it}(Y_{it}-F_{it})\omega_{i\tau}\tilde{\Psi}_{i\tau}\right)}{\sum_{t=1}^T \mathbb{E}_\phi(\omega_{it})} \right.$$
$$\left. -\frac{1}{2N}\sum_{i=1}^N \frac{\sum_{t=1}^T \mathbb{E}_\phi(\Psi_{it})\mathbb{E}_\phi(H_{it}\partial^2 F_{it}) + \mathbb{E}_\phi(\partial_{\pi^2}\Delta_{it})}{\sum_{t=1}^T \mathbb{E}_\phi(\omega_{it})} \right],$$

$$\overline{D}^\delta_\infty = \underset{N,T\to\infty}{\text{plim}} \left[-\frac{1}{2T}\sum_{t=1}^T \frac{\sum_{i=1}^N \mathbb{E}_\phi(\Psi_{it})\mathbb{E}_\phi(H_{it}\partial^2 F_{it})}{\sum_{i=1}^N \mathbb{E}_\phi(\omega_{it})} \right.$$
$$\left. +\frac{1}{2T}\sum_{t=1}^T \frac{\sum_{i=1}^N \mathbb{E}_\phi(\partial_{\pi^2}\Delta_{it})}{\sum_{i=1}^N \mathbb{E}_\phi(\omega_{it})} \right],$$

where $\tilde{\Psi}_{it}$ is the residual of the population regression of $-\partial_\pi \Delta_{it}/\mathbb{E}_\phi[\omega_{it}]$ on the space spanned by the incidental parameters under the metric given by $\mathbb{E}_\phi[\omega_{it}]$. If all the components of X_{it} are strictly exogenous, the first term of $\overline{B}^\delta_\infty$ is zero.

Example 2 (*Poisson Model*). Consider the partial effect

$$\Delta_{it}(\beta,\pi) = g_{it}(\beta)\exp(X'_{it}\beta+\pi),$$

where g_{it} does not depend on π. For example, $g_{it}(\beta) = \beta_k + \beta_j h(Z_{it})$ in (2.5). Using the notation previously introduced for this example, the components of the asymptotic bias are

$$\overline{B}^\delta_\infty = \underset{N,T\to\infty}{\text{plim}} \left[\frac{1}{N}\sum_{i=1}^N \frac{\sum_{t=1}^T \sum_{\tau=t+1}^T \mathbb{E}_\phi\left[(Y_{it}-\omega_{it})\omega_{i\tau}\tilde{g}_{i\tau}\right]}{\sum_{t=1}^T \mathbb{E}_\phi(\omega_{it})} \right],$$

and $\overline{D}^\delta_\infty = 0$, where $\tilde{g}_{it}$ is the residual of the population projection of g_{it} on the space spanned by the incidental parameters under a metric weighted by $\mathbb{E}_\phi[\omega_{it}]$. The asymptotic bias is zero if all the components of X_{it} are strictly exogenous or $g_{it}(\beta)$ is constant. The latter arises in the leading case of the partial effect of the kth component of X_{it} since $g_{it}(\beta) = \beta_k$. This no asymptotic bias result applies to any type of regressor, strictly exogenous or predetermined.

4.3. Bias corrected estimators

The results of the previous sections show that the asymptotic distributions of the fixed effects estimators of the model parameters and APEs can have biases of the same order as the variances under sequences where T grows at the same rate as N. This is the large-T version of the incidental parameter problem that invalidates any inference based on the fixed effect estimators even in large samples. In this section we describe how to construct analytical and jackknife bias corrections for the fixed effect estimators and give conditions for the asymptotic validity of these corrections.

The jackknife correction for the model parameter β in Eq. (3.4) is generic and applies to the panel model. For the APEs, the jackknife correction is formed similarly as

$$\widetilde{\delta}^J_{NT} = 3\widehat{\delta}_{NT} - \overline{\widetilde{\delta}}_{N,T/2} - \overline{\widetilde{\delta}}_{N/2,T},$$

where $\overline{\widetilde{\delta}}_{N,T/2}$ is the average of the 2 split jackknife estimators of the APE that use all the individuals and leave out the first and second halves of the time periods, and $\overline{\widetilde{\delta}}_{N/2,T}$ is the average of the 2 split jackknife estimators of the APE that use all the time periods and leave out half of the individuals.

The analytical corrections are constructed using sample analogs of the expressions in Theorems 4.1 and 4.2, replacing the true values of β and ϕ by the fixed effects estimators. To describe these corrections, we introduce some additional notation. For any function of the data, unobserved effects and parameters $g_{itj}(\beta, \alpha_i + \gamma_t, \alpha_i + \gamma_{t-j})$ with $0 \le j < t$, let $\widehat{g}_{itj} = g_{it}(\widehat{\beta}, \widehat{\alpha}_i + \widehat{\gamma}_t, \widehat{\alpha}_i + \widehat{\gamma}_{t-j})$ denote the fixed effects estimator, e.g., $\mathbb{E}_\phi\widehat{[\partial_{\pi^2}\ell_{it}]}$ denotes the fixed effects estimator of $\mathbb{E}_\phi[\partial_{\pi^2}\ell_{it}]$. Let $\widehat{\mathcal{H}}^{-1}_{(\alpha\alpha)}$, $\widehat{\mathcal{H}}^{-1}_{(\alpha\gamma)}$, $\widehat{\mathcal{H}}^{-1}_{(\gamma\alpha)}$, and $\widehat{\mathcal{H}}^{-1}_{(\gamma\gamma)}$ denote the blocks of the matrix $\widehat{\mathcal{H}}^{-1}$, where

$$\widehat{\mathcal{H}} = \begin{pmatrix} \widehat{\mathcal{H}}^*_{(\alpha\alpha)} & \widehat{\mathcal{H}}^*_{(\alpha\gamma)} \\ [\widehat{\mathcal{H}}^*_{(\alpha\gamma)}]' & \widehat{\mathcal{H}}^*_{(\gamma\gamma)} \end{pmatrix} + \frac{b}{\sqrt{NT}}\,vv',$$

$\widehat{\mathcal{H}}^*_{(\alpha\alpha)} = \text{diag}(-\sum_t \mathbb{E}_\phi\widehat{[\partial_{\pi^2}\ell_{it}]})/\sqrt{NT}$, $\widehat{\mathcal{H}}^*_{(\alpha\alpha)} = \text{diag}(-\sum_i \mathbb{E}_\phi\widehat{[\partial_{\pi^2}\ell_{it}]})/\sqrt{NT}$, and $\widehat{\mathcal{H}}^*_{(\alpha\gamma)it} = -\mathbb{E}_\phi\widehat{[\partial_{\pi^2}\ell_{it}]}/\sqrt{NT}$. Let

$$\widehat{\Xi}_{it} = -\frac{1}{\sqrt{NT}}\sum_{j=1}^N \sum_{\tau=1}^T \left(\widehat{\mathcal{H}}^{-1}_{(\alpha\alpha)ij} + \widehat{\mathcal{H}}^{-1}_{(\gamma\alpha)tj} \right.$$
$$\left. + \widehat{\mathcal{H}}^{-1}_{(\alpha\gamma)i\tau} + \widehat{\mathcal{H}}^{-1}_{(\gamma\gamma)t\tau} \right) \mathbb{E}_\phi\widehat{\left(\partial_{\beta\pi}\ell_{j\tau}\right)}.$$

The kth component of $\widehat{\Xi}_{it}$ corresponds to a least squares regression of $\mathbb{E}_\phi\widehat{\left(\partial_{\beta_k\pi}\ell_{it}\right)}/\mathbb{E}_\phi\widehat{(\partial_{\pi^2}\ell_{it})}$ on the space spanned by the incidental parameters weighted by $\mathbb{E}_\phi\widehat{(-\partial_{\pi^2}\ell_{it})}$.

The analytical bias corrected estimator of β^0 is

$$\widetilde{\beta}^A = \widehat{\beta} - \widehat{B}^\beta_{NT}/T - \widehat{D}^\beta_{NT}/N, \tag{4.6}$$

where $\widehat{B}^\beta_{NT} = \widehat{W}^{-1}\widehat{B}, \widehat{D}^\beta_{NT} = \widehat{W}^{-1}\widehat{D}$,

$$\widehat{B} = -\frac{1}{N}\sum_{i=1}^N \frac{\sum_{j=0}^L [T/(T-j)]\sum_{t=j+1}^T \mathbb{E}_\phi\left(\partial_\pi\ell_{i,t-j}\widehat{D_{\beta\pi}\ell_{it}}\right) + \frac{1}{2}\sum_{t=1}^T \mathbb{E}_\phi\widehat{(D_{\beta\pi^2}\ell_{it})}}{\sum_{t=1}^T \mathbb{E}_\phi\widehat{(\partial_{\pi^2}\ell_{it})}},$$

$$\widehat{D} = -\frac{1}{T}\sum_{t=1}^T \frac{\sum_{i=1}^N \left[\mathbb{E}_\phi\left(\widehat{\partial_\pi\ell_{it}D_{\beta\pi}\ell_{it}}\right) + \frac{1}{2}\mathbb{E}_\phi\widehat{(D_{\beta\pi^2}\ell_{it})} \right]}{\sum_{i=1}^N \mathbb{E}_\phi\widehat{(\partial_{\pi^2}\ell_{it})}},$$

$$\widehat{W} = -(NT)^{-1}\sum_{i=1}^N \sum_{t=1}^T \left[\mathbb{E}_\phi\widehat{(\partial_{\beta\beta'}\ell_{it})} - \mathbb{E}_\phi\left(\widehat{\partial_{\pi^2}\ell_{it}}\Xi_{it}\Xi'_{it}\right) \right], \tag{4.7}$$

and L is a trimming parameter for estimation of spectral expectations such that $L \to \infty$ and $L/T \to 0$ (Hahn and Kuersteiner, 2011). Here we use truncation instead of kernel smoothing in the estimation of spectral expectations following Hahn and Kuersteiner (2007). Note that, unlike for variance estimation, a kernel is not needed to ensure that the bias estimator be positive. Instead of choosing a value of L, our recommendation for practice is to conduct a sensitivity analysis by reporting estimates for multiple values of L starting from $L = 0$. From our experience based on extensive Monte Carlo simulations, we do not recommend values of L greater than 4, because the finite-sample dispersion of the estimator quickly increases with L. We refer to Section 5 for an example of sensitivity analysis with respect to L. The factor $T/(T-j)$ is a degrees of freedom adjustment that rescales the time series averages $T^{-1}\sum_{t=j+1}^T$ by the number of observations instead of by T. Similar corrections for conditional mean models can be formed using the sample analogs of the

expressions of $\overline{B}_\infty$ and $\overline{D}_\infty$ in Remark 3. We do not spell out these estimators for the sake of brevity.

Asymptotic $(1-p)$–confidence intervals for the components of β can be formed as

$$\widetilde{\beta}_k^A \pm z_{1-p}\sqrt{\widehat{W}_{kk}^{-1}/(NT)}, \quad k=\{1,\ldots,\dim\beta\},$$

where z_{1-p} is the $(1-p)$–quantile of the standard normal distribution, and $\widehat{W}_{kk}^{-1}$ is the (k,k)-element of the matrix $\widehat{W}^{-1}$. In conditional moment models we replace $\widehat{W}_{kk}$ by the (k,k)-element of the matrix $\widehat{W}^{-1}\widehat{\Omega}\widehat{W}^{-1}$, where

$$\widehat{\Omega} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\tau=1}^{T}\mathbb{E}_\phi\left[D_\beta\widehat{\ell_{it}(D_\beta\ell_{i\tau})'}\right].$$

We have implemented the analytical correction at the level of the estimator. Alternatively, we can implement the correction at the level of the score or first order conditions by solving

$$(NT)^{-1/2}\partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta)) = \widehat{B}/T + \widehat{D}/N, \tag{4.8}$$

for β. Global concavity of the objective function guarantees that the solution to (4.8) is unique. Other possible extensions such as continuously updated score corrections where $\overline{B}_\infty$ and $\overline{D}_\infty$ are estimated together with β, corrections at the level of the objective function, or iterative corrections are left to future research.

The analytical bias corrected estimator of δ_{NT}^0 is

$$\widetilde{\delta}^A = \widetilde{\delta} - \widehat{B}^\delta/T - \widehat{D}^\delta/N,$$

where $\widetilde{\delta}$ is the APE constructed from a bias corrected estimator of δ_{NT}^0. Let

$$\widehat{\Psi}_{it} = -\frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\sum_{\tau=1}^{T}\left(\widehat{\mathcal{H}}_{(\alpha\alpha)ij}^{-1} + \widehat{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}\right.$$
$$\left. + \widehat{\mathcal{H}}_{(\alpha\gamma)i\tau}^{-1} + \widehat{\mathcal{H}}_{(\gamma\gamma)t\tau}^{-1}\right)\widehat{\partial_\pi\Delta_{j\tau}}.$$

The fixed effects estimators of the components of the asymptotic bias are

$$\widehat{B}^\delta = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j=0}^{L}[T/(T-j)]\sum_{t=j+1}^{T}\mathbb{E}_\phi\left(\partial_\pi\widehat{\ell_{i,t-j}\partial_{\pi^2}\ell_{it}\Psi_{it}}\right)}{\sum_{t=1}^{T}\mathbb{E}_\phi\widehat{\left(\partial_{\pi^2}\ell_{it}\right)}}$$
$$- \frac{1}{2N}\sum_{i=1}^{N}\frac{\sum_{t=1}^{T}\left[\mathbb{E}_\phi\widehat{(\partial_{\pi^2}\Delta_{it})} - \mathbb{E}_\phi\widehat{(\partial_{\pi^3}\ell_{it})}\mathbb{E}_\phi\widehat{(\Psi_{it})}\right]}{\sum_{t=1}^{T}\mathbb{E}_\phi\widehat{\left(\partial_{\pi^2}\ell_{it}\right)}},$$

$$\widehat{D}^\delta = \frac{1}{T}\sum_{t=1}^{T}\frac{\sum_{i=1}^{N}\left[\mathbb{E}_\phi\left(\partial_\pi\widehat{\ell_{it}\partial_{\pi^2}\ell_{it}\Psi_{it}}\right) - \frac{1}{2}\mathbb{E}_\phi\widehat{(\partial_{\pi^2}\Delta_{it})} + \frac{1}{2}\mathbb{E}_\phi\widehat{(\partial_{\pi^3}\ell_{it})}\mathbb{E}_\phi\widehat{(\Psi_{it})}\right]}{\sum_{i=1}^{N}\mathbb{E}_\phi\widehat{\left(\partial_{\pi^2}\ell_{it}\right)}}.$$

The estimator of the asymptotic variance depends on the sampling properties of the unobserved effects. Under the independence assumption of Remark 4,

$$\widehat{V}^\delta = \frac{r_{NT}^2}{N^2T^2}\sum_{i=1}^{N}\left[\sum_{t,\tau=1}^{T}\widehat{\overline{\Delta}}_{it}\widehat{\overline{\Delta}}_{i\tau}' + \sum_{t=1}^{T}\sum_{j\neq i}\widehat{\overline{\Delta}}_{it}\widehat{\overline{\Delta}}_{jt}'\right.$$
$$\left. + \sum_{t=1}^{T}\mathbb{E}_\phi\widehat{\left(\Gamma_{it}\Gamma_{it}'\right)}\right], \tag{4.9}$$

where $\widehat{\overline{\Delta}}_{it} = \widehat{\Delta}_{it} - N^{-1}\sum_{i=1}^{N}\widehat{\Delta}_{it}$ under identical distribution over i, $\widehat{\overline{\Delta}}_{it} = \widehat{\Delta}_{it} - T^{-1}\sum_{t=1}^{T}\widehat{\Delta}_{it}$ under stationarity over t, and $\widehat{\overline{\Delta}}_{it} = \widehat{\Delta}_{it} - \widehat{\delta}$ under both. Note that we do not need to specify

the convergence rate r_{NT} to make inference because the standard errors $\sqrt{\widehat{V}^\delta}/r_{NT}$ do not depend on r_{NT}. Bias corrected estimators and confidence intervals can be constructed in the same fashion as for the model parameter.

We use the following homogeneity assumption to show the validity of the jackknife corrections for the model parameters and APEs. It implies that $\widetilde{\beta}_{N,T/2} - \widehat{\beta}_{NT} = \overline{B}_\infty^\beta/T + o_P(T^{-1}\vee N^{-1})$ and $\widetilde{\beta}_{N/2,T} - \widehat{\beta}_{NT} = \overline{D}_\infty^\beta/N + o_P(T^{-1}\vee N^{-1})$, which are weaker but higher level sufficient conditions for the validity of the jackknife for the model parameter. For APEs, Assumption 4.3 also ensures that these effects do not change with T and N, i.e. $\delta_{NT}^0 = \delta^0$. The analytical corrections *do not* require this assumption.

Assumption 4.3 (*Unconditional Homogeneity*)**.** The sequence $\{(Y_{it}, X_{it}, \alpha_i, \gamma_t) : 1 \leq i \leq N, 1 \leq t \leq T\}$ is identically distributed across i and strictly stationary across t, for each N, T.

This assumption might seem restrictive for dynamic models where X_{it} includes lags of the dependent variable because in this case it restricts the unconditional distribution of the initial conditions of Y_{it}. Note, however, that Assumption 4.3 allows the initial conditions to depend on the unobserved effects. In other words, it does not impose that the initial conditions are generated from the stationary distribution of Y_{it} conditional on X_i^t and ϕ. Assumption 4.3 rules out time trends and structural breaks in the processes for the unobserved effects and observed variables.

Remark 6 (*Test of Homogeneity*)**.** Assumption 4.3 is a sufficient condition for the validity of the jackknife corrections. It has the testable implications that the probability limits of the fixed effects estimator are the same in all the partitions of the panel. For example, it implies that $\beta_{N,T/2}^1 = \beta_{N,T/2}^2$, where $\beta_{N,T/2}^1$ and $\beta_{N,T/2}^2$ are the probability limits of the fixed effects estimators of β in the subpanels that include all the individuals and the first and second halves of the time periods, respectively. These implications can be tested using variations of the Chow-type test proposed in Dhaene and Jochmans (2015). We provide an example of the application of these tests to our setting in Section S.1.1 of the supplemental material.

The following theorems are the main results of this section. They show that the analytical and jackknife bias corrections eliminate the bias from the asymptotic distribution of the fixed effects estimators of the model parameters and APEs without increasing variance, and that the estimators of the asymptotic variances are consistent.

Theorem 4.3 (*Bias Corrections for $\widehat{\beta}$*)**.** *Under the conditions of Theorem 4.1,*

$$\widehat{W} \to_P \overline{W}_\infty,$$

and, if $L \to \infty$ and $L/T \to 0$,

$$\sqrt{NT}(\widetilde{\beta}^A - \beta^0) \to_d \mathcal{N}(0, \overline{W}_\infty^{-1}).$$

Under the conditions of Theorem 4.1 and Assumption 4.3,

$$\sqrt{NT}(\widetilde{\beta}^J - \beta^0) \to_d \mathcal{N}(0, \overline{W}_\infty^{-1}).$$

Theorem 4.4 (*Bias Corrections for $\widehat{\delta}$*)**.** *Under the conditions of Theorems 4.1 and 4.2,*

$$\widehat{V}^\delta \to_P \overline{V}_\infty^\delta,$$

and, if $L \to \infty$ and $L/T \to 0$,

$$r_{NT}(\widetilde{\delta}^A - \delta_{NT}^0) \to_d \mathcal{N}(0, \overline{V}_\infty^\delta).$$

Under the conditions of Theorems *4.1 and* 4.2, *and* Assumption *4.3,*

$$r_{NT}(\widetilde{\delta}^J - \delta^0) \to_d \mathcal{N}(0, \overline{V}^\delta_\infty).$$

Remark 7 (*Rate of Convergence*). The rate of convergence r_{NT} depends on the properties of the sampling process for the explanatory variables and unobserved effects (see Remark 4).

5. Monte Carlo experiments

This section reports evidence on the finite sample behavior of fixed effects estimators of model parameters and APEs in static models with strictly exogenous regressors and dynamic models with predetermined regressors such as lags of the dependent variable. We analyze the performance of uncorrected and bias-corrected fixed effects estimators in terms of bias and inference accuracy of their asymptotic distribution. In particular we compute the biases, standard deviations, and root mean squared errors of the estimators, the ratio of average standard errors to the simulation standard deviations (SE/SD); and the empirical coverages of confidence intervals with 95% nominal value (p; .95).[14] Overall, we find that the analytically corrected estimators dominate the uncorrected and jackknife corrected estimators.[15] A possible explanation for the better finite-sample performance of the analytical over the jackknife corrections is that the jackknife increases dispersion because the components of the bias are estimated from subsamples that include half of the observations of the panel. We observe this variance increase in all our numerical examples, specially in short panels. The jackknife corrections are also more sensitive than the analytical corrections to Assumption 4.3. All the results are based on 500 replications. The designs correspond to static and dynamic probit models. As in the analytical example of Section 3.3, we find that our large T asymptotic approximations capture well the behavior of the fixed effects estimator and the bias corrections in moderately long panels with $N = 56$ and $T = 14$.

5.1. Static probit model

The data generating process is

$$Y_{it} = \mathbf{1}\{X_{it}\beta + \alpha_i + \gamma_t > \varepsilon_{it}\}, \quad (i = 1, \ldots, N; \ t = 1, \ldots, T),$$

where $\alpha_i \sim \mathcal{N}(0, 1/16)$, $\gamma_t \sim \mathcal{N}(0, 1/16)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, and $\beta = 1$. We consider two alternative designs for X_{it}: autoregressive process and linear trend process both with individual and time effects. In the first design, $X_{it} = X_{i,t-1}/2 + \alpha_i + \gamma_t + \upsilon_{it}$, $\upsilon_{it} \sim \mathcal{N}(0, 1/2)$, and $X_{i0} \sim \mathcal{N}(0, 1)$. In the second design, $X_{it} = 2t/T + \alpha_i + \gamma_t + \upsilon_{it}$, $\upsilon_{it} \sim \mathcal{N}(0, 3/4)$, which violates Assumption 4.3. In both designs X_{it} is strictly exogenous with respect to ε_{it} conditional on the individual and time effects. The variables α_i, γ_t, ε_{it}, υ_{it}, and X_{i0} are independent and *i.i.d.* across individuals and time periods. We generate panel data sets with $N = 56$ individuals and three different numbers of time periods T: 14, 28 and 56.[16]

Table 3 reports the results for the probit coefficient β, and the APE of X_{it}. We compute the APE using (2.4). Throughout the table, MLE-FETE corresponds to the probit maximum likelihood

estimator with individual and time fixed effects, Analytical is the bias corrected estimator that uses the analytical correction, and Jackknife is the bias corrected estimator that uses SPJ in both the individual and time dimensions. The cross-sectional division in the jackknife follows the order of the observations. All the results are reported in percentage of the true parameter value.

We find that the bias is of the same order of magnitude as the standard deviation for the uncorrected estimator of the probit coefficient causing severe undercoverage of the confidence intervals. This result holds for both designs and all the sample sizes considered. The bias corrections, specially Analytical, remove the bias without increasing dispersion, and produce substantial improvements in rmse and coverage probabilities. For example, Analytical reduces rmse by 50% and increases coverage by 26% in the first design with $T = 14$. As in Hahn and Newey (2004) and Fernández-Val (2009), we find very little bias in the uncorrected estimates of the APE, despite the large bias in the probit coefficients. Jackknife performs relatively worse in the second design that does not satisfy Assumption 4.3.

5.2. Dynamic probit model

The data generating process is

$$Y_{it} = \mathbf{1}\{Y_{i,t-1}\beta_Y + Z_{it}\beta_Z + \alpha_i + \gamma_t > \varepsilon_{it}\},$$
$$(i = 1, \ldots, N; \ t = 1, \ldots, T),$$
$$Y_{i0} = \mathbf{1}\{Z_{i0}\beta_Z + \alpha_i + \gamma_0 > \varepsilon_{i0}\},$$

where $\alpha_i \sim \mathcal{N}(0, 1/16)$, $\gamma_t \sim \mathcal{N}(0, 1/16)$, $\varepsilon_{it} \sim \mathcal{N}(0, 1)$, $\beta_Y = 0.5$, and $\beta_Z = 1$. We consider two alternative designs for Z_{it}: autoregressive process and linear trend process both with individual and time effects. In the first design, $Z_{it} = Z_{i,t-1}/2 + \alpha_i + \gamma_t + \upsilon_{it}$, $\upsilon_{it} \sim \mathcal{N}(0, 1/2)$, and $Z_{i0} \sim \mathcal{N}(0, 1)$. In the second design, $Z_{it} = 1.5t/T + \alpha_i + \gamma_t + \upsilon_{it}$, $\upsilon_{it} \sim \mathcal{N}(0, 3/4)$, which violates Assumption 4.3. The variables α_i, γ_t, ε_{it}, υ_{it}, and Z_{i0} are independent and *i.i.d.* across individuals and time periods. We generate panel data sets with $N = 56$ individuals and three different numbers of time periods T: 14, 28 and 56.

Table 4 reports the simulation results for the probit coefficient β_Y and the APE of $Y_{i,t-1}$. We compute the partial effect of $Y_{i,t-1}$ using the expression in Eq. (2.3) with $X_{it,k} = Y_{i,t-1}$. This effect is commonly reported as a measure of state dependence for dynamic binary processes. Table 5 reports the simulation results for the estimators of the probit coefficient β_Z and the APE of Z_{it}. We compute the partial effect using (2.4) with $X_{it,k} = Z_{it}$. Throughout the tables, we compare the same estimators as for the static model. For the analytical correction we consider two versions, Analytical ($L = 1$) sets the trimming parameter to estimate spectral expectations L to one, whereas Analytical ($L = 2$) sets L to two.[17] Again, all the results in the tables are reported in percentage of the true parameter value.

The results in Table 4 show important biases toward zero for *both* the probit coefficient and the APE of $Y_{i,t-1}$ in the two designs. This bias can indeed be substantially larger than the corresponding standard deviation for short panels yielding coverage probabilities below 70% for $T = 14$. The analytical corrections significantly reduce biases and rmse, bring coverage probabilities close to their nominal level, and have little sensitivity to the trimming parameter L. The jackknife corrections reduce bias but increase dispersion, producing less drastic improvements in rmse and coverage than the analytical corrections. The results for the APE of Z_{it} in Table 5

---

[14] The standard errors are computed using the expressions (4.7) and (4.9) with $\widehat{\Delta}_{it} = \widehat{\Delta}_{it} - \widehat{\delta}$, evaluated at uncorrected estimates of the parameters. We find little difference in performance of constructing standard errors based on corrected estimates.

[15] Kristensen and Salanié (2013) also found that analytical corrections dominate jackknife corrections to reduce the bias of approximate estimators.

[16] Following a suggestion from an anonymous referee, we obtained results for panel data sets with $T = 56$ and N in {14, 28, 56}. These results are similar to the results reported and are available from the authors upon request.

[17] In results not reported for brevity, we find little difference in performance of increasing the trimming parameters to $L = 3$ and $L = 4$. These results are available from the authors upon request.

Table 3
Finite-sample properties in static probit model ($N = 56$).

	Coefficient					APE				
	Bias	Std. Dev.	RMSE	SE/SD	p; .95	Bias	Std. Dev.	RMSE	SE/SD	p; .95
Design 1: *autoregressive with individual and time effects*										
$T = 14$										
MLE-FETE	14	11	18	0.88	0.71	1	8	8	0.93	0.93
Analytical	1	9	9	1.05	0.97	0	8	8	0.94	0.94
Jackknife	−6	11	13	0.88	0.87	1	9	9	0.80	0.87
$T = 28$										
MLE-FETE	7	7	10	0.93	0.81	1	5	5	1.02	0.96
Analytical	0	6	6	1.03	0.95	0	5	5	1.03	0.96
Jackknife	−2	7	7	0.96	0.92	0	6	6	0.94	0.94
$T = 56$										
MLE-FETE	5	4	6	1.01	0.82	0	4	4	0.98	0.94
Analytical	0	4	4	1.07	0.98	0	4	4	0.98	0.94
Jackknife	−1	4	4	1.00	0.95	0	4	4	0.93	0.93
Design 2: *linear trend with individual and time effects*										
$T = 14$										
MLE-FETE	18	13	22	0.85	0.62	−3	10	10	0.76	0.85
Analytical	0	10	10	1.10	0.96	−4	10	10	0.77	0.84
Jackknife	−13	20	23	0.55	0.74	−3	11	11	0.71	0.82
$T = 28$										
MLE-FETE	8	7	11	0.93	0.79	−2	7	7	0.84	0.88
Analytical	0	7	7	1.05	0.95	−2	7	7	0.84	0.89
Jackknife	−5	8	9	0.91	0.87	−2	7	7	0.81	0.87
$T = 56$										
MLE-FETE	6	5	7	0.97	0.77	0	5	5	0.85	0.89
Analytical	0	4	4	1.05	0.97	0	5	5	0.85	0.89
Jackknife	−2	5	5	0.99	0.92	0	5	5	0.83	0.89

Notes: All the entries are in percentage of the true parameter value. 500 repetitions. Data generated from the probit model: $Y_{it} = 1(\beta X_{it} + \alpha_i + \gamma_t > \varepsilon_{it})$, with $\varepsilon_{it} \sim$ i.i.d. $N(0, 1)$, $\alpha_i \sim$ i.i.d. $N(0, 1/16)$, $\gamma_t \sim$ i.i.d. $N(0, 1/16)$ and $\beta = 1$. In design 1, $X_{it} = X_{i,t-1}/2 + \alpha_i + \gamma_t + v_{it}$, $v_{it} \sim$ i.i.d. $N(0, 1/2)$, and $X_{i0} \sim N(0, 1)$. In design 2, $X_{it} = 2t/T + \alpha_i + \gamma_t + v_{it}$, and $v_{it} \sim$ i.i.d. $N(0, 3/4)$, independent of $\alpha_i \gamma_t$. Average effect is $\beta E[\varphi(\beta X_{it} + \alpha_i + \gamma_t)]$, where $\varphi()$ is the PDF of the standard normal distribution. MLE-FETE is the probit maximum likelihood estimator with individual and time fixed effects; Analytical is the bias corrected estimator that uses an analytical correction; and Jackknife is the bias corrected estimator that uses split panel jackknife in both the individual and time dimension.

are similar to the static probit model. There are significant bias and undercoverage of confidence intervals for the coefficient β_Z, which are removed by the corrections, whereas there are little bias and undercoverage in the APE. As in the static model, Jackknife performs relatively worse in the second design.

6. Concluding remarks

In this paper we develop analytical and jackknife corrections for fixed effects estimators of model parameters and APEs in semiparametric nonlinear panel models with additive individual and time effects. Our analysis applies to conditional maximum likelihood estimators with concave log-likelihood functions, and therefore covers logit, probit, ordered probit, ordered logit, Poisson, negative binomial, and Tobit estimators, which are the most popular nonlinear estimators in empirical economics.

We are currently developing similar corrections for nonlinear models with interactive individual and time effects (Chen et al., 2014). Another interesting avenue of future research is to derive higher-order expansions for fixed effects estimators with individual and time effects. These expansions are needed to justify theoretically the validity of alternative corrections based on the leave-one-observation-out panel jackknife method of Hahn and Newey (2004).

Appendix A. Notation and choice of norms

We write A' for the transpose of a matrix or vector A. We use $\mathbb{1}_n$ for the $n \times n$ identity matrix, and 1_n for the column vector of length n whose entries are all unity. For square $n \times n$ matrices B, C, we use $B > C$ (or $B \geq C$) to indicate that $B - C$ is positive (semi) definite.

We write wpa1 for "with probability approaching one" and wrt for "with respect to". All the limits are taken as $N, T \to \infty$ jointly.

As in the main text, we usually suppress the dependence on NT of all the sequences of functions and parameters to lighten the notation, e.g. we write $\mathcal{L}$ for $\mathcal{L}_{NT}$ and ϕ for ϕ_{NT}. Let

$$\mathcal{S}(\beta, \phi) = \partial_\phi \mathcal{L}(\beta, \phi), \qquad \mathcal{H}(\beta, \phi) = -\partial_{\phi\phi'} \mathcal{L}(\beta, \phi),$$

where $\partial_x f$ denotes the partial derivative of f with respect to x, and additional subscripts denote higher-order partial derivatives. We refer to the dim ϕ-vector $\mathcal{S}(\beta, \phi)$ as the incidental parameter score, and to the dim $\phi \times$ dim ϕ matrix $\mathcal{H}(\beta, \phi)$ as the incidental parameter Hessian. We omit the arguments of the functions when they are evaluated at the true parameter values (β^0, ϕ^0), e.g. $\mathcal{H} = \mathcal{H}(\beta^0, \phi^0)$. We use a bar to indicate expectations conditional on ϕ, e.g. $\partial_\beta \overline{\mathcal{L}} = \mathbb{E}_\phi[\partial_\beta \mathcal{L}]$, and a tilde to denote variables in deviations with respect to expectations, e.g. $\partial_\beta \widetilde{\mathcal{L}} = \partial_\beta \mathcal{L} - \partial_\beta \overline{\mathcal{L}}$.

We use the Euclidean norm $\|.\|$ for vectors of dimension dim β, and we use the norm induced by the Euclidean norm for the corresponding matrices and tensors, which we also denote by $\|.\|$. For matrices of dimension dim $\beta \times$ dim β this induced norm is the spectral norm. The generalization of the spectral norm to higher order tensors is straightforward, e.g. the induced norm of the dim $\beta \times$ dim $\beta \times$ dim β tensor of third partial derivatives of $\mathcal{L}(\beta, \phi)$ wrt β is given by

$$\left\| \partial_{\beta\beta\beta} \mathcal{L}(\beta, \phi) \right\| = \max_{\{u, v \in \mathbb{R}^{\dim \beta}, \|u\|=1, \|v\|=1\}}$$
$$\times \left\| \sum_{k,l=1}^{\dim \beta} u_k v_l \partial_{\beta\beta_k\beta_l} \mathcal{L}(\beta, \phi) \right\|.$$

This choice of norm is immaterial for the asymptotic analysis because dim β is fixed with the sample size.

Table 4
Finite-sample properties in dynamic probit model: lagged dependent variable ($N = 56$).

	Coefficient of $Y_{i,t-1}$					APE of $Y_{i,t-1}$				
	Bias	Std. Dev.	RMSE	SE/SD	p; .95	Bias	Std. Dev.	RMSE	SE/SD	p; .95
Design 1: autoregressive with individual and time effects										
$T = 14$										
MLE-FETE	−43	29	52	0.94	0.64	−51	26	57	0.93	0.43
Analytical ($L = 1$)	−4	26	26	1.07	0.96	−4	27	28	0.88	0.92
Analytical ($L = 2$)	−3	27	27	1.01	0.95	−3	29	29	0.83	0.91
Jackknife	12	32	34	0.86	0.89	−3	33	33	0.73	0.84
$T = 28$										
MLE-FETE	−20	19	28	0.96	0.80	−27	18	32	0.96	0.67
Analytical ($L = 1$)	−3	18	18	1.04	0.95	−3	19	19	0.94	0.92
Analytical ($L = 2$)	−1	18	18	1.01	0.94	0	19	19	0.91	0.91
Jackknife	2	19	19	0.96	0.94	−1	21	21	0.86	0.90
$T = 56$										
MLE-FETE	−9	13	16	0.99	0.89	−13	13	19	0.98	0.83
Analytical ($L = 1$)	−1	12	12	1.03	0.95	−1	13	13	0.97	0.94
Analytical ($L = 2$)	0	13	13	1.02	0.95	0	14	13	0.96	0.93
Jackknife	1	13	13	0.99	0.94	1	14	14	0.92	0.93
Design 2: linear trend with individual and time effects										
$T = 14$										
MLE-FETE	−48	35	60	0.94	0.69	−59	28	65	0.95	0.43
Analytical ($L = 1$)	−6	30	31	1.11	0.97	−11	30	32	0.88	0.88
Analytical ($L = 2$)	−6	32	33	1.05	0.95	−11	32	34	0.83	0.87
Jackknife	8	46	47	0.73	0.86	−19	35	40	0.75	0.80
$T = 28$										
MLE-FETE	−23	24	33	0.94	0.79	−33	22	39	0.93	0.61
Analytical ($L = 1$)	−4	22	22	1.03	0.97	−6	22	23	0.90	0.90
Analytical ($L = 2$)	−2	22	22	1.00	0.95	−5	23	24	0.87	0.90
Jackknife	1	25	25	0.87	0.91	−7	25	26	0.81	0.87
$T = 56$										
MLE-FETE	−11	15	19	1.01	0.89	−17	15	23	0.97	0.78
Analytical ($L = 1$)	−3	14	15	1.07	0.95	−3	16	16	0.96	0.94
Analytical ($L = 2$)	−2	14	15	1.05	0.95	−2	16	16	0.95	0.94
Jackknife	−1	16	16	0.95	0.94	−3	17	17	0.88	0.91

Notes: All the entries are in percentage of the true parameter value. 500 repetitions. Data generated from the probit model: $Y_{it} = 1(\beta_Y Y_{i,t-1} + \beta_Z Z_{it} + \alpha_i + \gamma_t > \varepsilon_{it})$, with $Y_{i0} = 1(\beta_Z Z_{i0} + \alpha_i + \gamma_0 > \varepsilon_{i0})$, $\varepsilon_{it} \sim$ i.i.d. $N(0, 1)$, $\alpha_i \sim$ i.i.d. $N(0, 1/16)$, $\gamma_t \sim$ i.i.d. $N(0, 1/16)$, $\beta_Y = 0.5$, and $\beta_Z = 1$. In design 1, $Z_{it} = Z_{i,t-1}/2 + \alpha_i + \gamma_t + v_{it}$, $v_{it} \sim$ i.i.d. $N(0, 1/2)$, and $Z_{i0} \sim N(0, 1)$. In design 2, $Z_{it} = 1.5t/T + v_{it}$, and $v_{it} \sim$ i.i.d. $N(0, 3/4)$, independent of $\alpha_i \gamma_t$. Average effect is $E[\Phi(\beta_Y + \beta_Z Z_{it} + \alpha_i + \gamma_t) - \Phi(\beta_Z Z_{it} + \alpha_i + \gamma_t)]$, where $\Phi()$ is the CDF of the standard normal distribution. MLE-FETE is the probit maximum likelihood estimator with individual and time fixed effects; Analytical ($L = l$) is the bias corrected estimator that uses an analytical correction with l lags to estimate the spectral expectations; and Jackknife is the bias corrected estimator that uses split panel jackknife in both the individual and time dimension.

In contrast, it is important what norms we choose for vectors of dimension $\dim \phi$, and their corresponding matrices and tensors, because $\dim \phi$ is increasing with the sample size. For vectors of dimension $\dim \phi$, we use the ℓ_q-norm

$$\|\phi\|_q = \left(\sum_{g=1}^{\dim \phi} |\phi_g|^q \right)^{1/q},$$

where $2 \leq q \leq \infty$.[18] The particular value $q = 8$ will be chosen later.[19] We use the norms that are induced by the ℓ_q-norm for the corresponding matrices and tensors, e.g. the induced q-norm of the $\dim \phi \times \dim \phi \times \dim \phi$ tensor of third partial derivatives of $\mathcal{L}(\beta, \phi)$ wrt ϕ is

$$\left\| \partial_{\phi\phi\phi} \mathcal{L}(\beta, \phi) \right\|_q = \max_{\{u, v \in \mathbb{R}^{\dim \phi}, \, \|u\|_q = 1, \, \|v\|_q = 1\}}$$

$$\times \left\| \sum_{g,h=1}^{\dim \phi} u_g \, v_h \, \partial_{\phi\phi_g \phi_h} \mathcal{L}(\beta, \phi) \right\|_q. \tag{A.1}$$

Note that in general the ordering of the indices of the tensor would matter in the definition of this norm, with the first index having a special role. However, since partial derivatives like $\partial_{\phi_g \phi_h \phi_l} \mathcal{L}(\beta, \phi)$ are fully symmetric in the indices g, h, l, the ordering is not important in their case.

For mixed partial derivatives of $\mathcal{L}(\beta, \phi)$ wrt β and ϕ, we use the norm that is induced by the Euclidean norm on $\dim \beta$-vectors and the q-norm on $\dim \phi$-indices, e.g.

$$\left\| \partial_{\beta\beta\phi\phi\phi} \mathcal{L}(\beta, \phi) \right\|_q = \max_{\{u, v \in \mathbb{R}^{\dim \beta}, \, \|u\| = 1, \, \|v\| = 1\}}$$

$$\times \max_{\{w, x \in \mathbb{R}^{\dim \phi}, \, \|w\|_q = 1, \, \|x\|_q = 1\}}$$

$$\left\| \sum_{k,l=1}^{\dim \beta} \sum_{g,h=1}^{\dim \phi} u_k \, v_l \, w_g \, x_h \, \partial_{\beta_k \beta_l \phi \phi_g \phi_h} \mathcal{L}(\beta, \phi) \right\|_q, \tag{A.2}$$

where we continue to use the notation $\|.\|_q$, even though this is a mixed norm.

Note that for $w, x \in \mathbb{R}^{\dim \phi}$ and $q \geq 2$,

$$|w'x| \leq \|w\|_q \|x\|_{q/(q-1)} \leq (\dim \phi)^{(q-2)/q} \|w\|_q \|x\|_q.$$

---

[18] We use the letter q instead of p to avoid confusion with the use of p for probability.

[19] The main reason not to choose $q = \infty$ is the assumption $\|\widetilde{\mathcal{H}}\|_q = o_P(1)$ below, which is used to guarantee that $\|\mathcal{H}^{-1}\|_q$ is of the same order as $\|\overline{\mathcal{H}}^{-1}\|_q$. If we assume $\|\mathcal{H}^{-1}\|_q = \mathcal{O}_P(1)$ directly instead of $\|\overline{\mathcal{H}}^{-1}\|_q = \mathcal{O}_P(1)$, then we can set $q = \infty$.

Table 5
Finite-sample properties in dynamic probit model: exogenous regressor ($N = 56$).

	Coefficient of Z_{it}					APE of Z_{it}				
	Bias	Std. Dev.	RMSE	SE/SD	p; .95	Bias	Std. Dev.	RMSE	SE/SD	p; .95
Design 1: autoregressive with individual and time effects										
	T = 14									
MLE-FETE	19	12	22	0.91	0.57	5	9	10	0.90	0.88
Analytical ($L = 1$)	2	10	10	1.11	0.97	2	9	9	0.91	0.92
Analytical ($L = 2$)	2	10	10	1.10	0.97	2	9	9	0.91	0.92
Jackknife	−7	12	14	0.88	0.85	4	10	11	0.80	0.84
	T = 28									
MLE-FETE	10	7	12	0.94	0.69	3	6	7	0.94	0.92
Analytical ($L = 1$)	1	7	7	1.05	0.97	1	6	6	0.95	0.95
Analytical ($L = 2$)	1	7	7	1.04	0.97	1	6	6	0.95	0.94
Jackknife	−2	7	8	0.94	0.91	1	7	7	0.88	0.91
	T = 56									
MLE-FETE	6	5	8	0.98	0.77	1	5	5	0.95	0.92
Analytical ($L = 1$)	0	4	4	1.05	0.97	0	5	5	0.96	0.94
Analytical ($L = 2$)	0	4	4	1.05	0.97	0	5	5	0.96	0.94
Jackknife	−1	5	5	0.98	0.95	0	5	5	0.92	0.91
Design 2: linear trend with individual and time effects										
	T = 14									
MLE-FETE	19	13	23	0.86	0.62	−3	11	11	0.75	0.83
Analytical ($L = 1$)	1	10	10	1.13	0.97	−3	11	11	0.76	0.83
Analytical ($L = 2$)	1	10	10	1.12	0.97	−4	11	11	0.75	0.82
Jackknife	−16	22	27	0.52	0.70	−2	12	12	0.67	0.80
	T = 28									
MLE-FETE	10	8	13	0.88	0.72	−2	8	8	0.83	0.86
Analytical ($L = 1$)	1	7	7	1.01	0.96	−2	8	8	0.83	0.87
Analytical ($L = 2$)	1	7	7	1.01	0.96	−2	8	8	0.83	0.87
Jackknife	−4	8	9	0.90	0.87	−2	8	8	0.79	0.86
	T = 56									
MLE-FETE	6	5	8	0.89	0.73	−1	6	6	0.83	0.90
Analytical ($L = 1$)	0	5	5	0.98	0.95	−1	6	6	0.84	0.90
Analytical ($L = 2$)	0	5	5	0.98	0.95	−1	6	6	0.83	0.90
Jackknife	−2	5	5	0.94	0.92	−1	6	6	0.83	0.89

Notes: All the entries are in percentage of the true parameter value. 500 repetitions. Data generated from the probit model: $Y_{it} = 1(\beta_Y Y_{i,t-1} + \beta_Z Z_{it} + \alpha_i + \gamma_t > \varepsilon_{it})$, with $Y_{i0} = 1(\beta_Z Z_{i0} + \alpha_i + \gamma_0 > \varepsilon_{i0})$, $\varepsilon_{it} \sim$ i.i.d. $N(0, 1)$, $\alpha_i \sim$ i.i.d. $N(0, 1/16)$, $\gamma_t \sim$ i.i.d. $N(0, 1/16)$, $\beta_Y = 0.5$, and $\beta_Z = 1$. In design 1, $Z_{it} = Z_{i,t-1}/2 + \alpha_i + \gamma_t + v_{it}$, $v_{it} \sim$ i.i.d. $N(0, 1/2)$, and $Z_{i0} \sim N(0, 1)$. In design 2, $Z_{it} = 1.5t/T + v_{it}$, and $v_{it} \sim$ i.i.d. $N(0, 3/4)$, independent of $\alpha_i y \gamma_t$. Average effect is $\beta_Z E[\varphi(\beta_Y Y_{i,t-1} + \beta_Z Z_{it} + \alpha_i + \gamma_t)]$, where $\varphi()$ is the PDF of the standard normal distribution. MLE-FETE is the probit maximum likelihood estimator with individual and time fixed effects; Analytical ($L = l$) is the bias corrected estimator that uses an analytical correction with l lags to estimate the spectral expectations; and Jackknife is the bias corrected estimator that uses split panel jackknife in both the individual and time dimension.

Thus, whenever we bound a scalar product of vectors, matrices and tensors in terms of the above norms we have to account for this additional factor $(\dim \phi)^{(q-2)/q}$. For example,

$$\left| \sum_{k,l=1}^{\dim \beta} \sum_{f,g,h=1}^{\dim \phi} u_k v_l w_f x_h y_f \, \partial_{\beta_k \beta_l \phi_f \phi_g \phi_h} \mathcal{L}(\beta, \phi) \right|$$
$$\leq (\dim \phi)^{(q-2)/q} \|u\| \|v\| \|w\|_q \|x\|_q \|y\|_q \, \left\| \partial_{\beta\beta\phi\phi\phi} \mathcal{L}(\beta, \phi) \right\|_q .$$

For higher-order tensors, we use the notation $\partial_{\phi\phi\phi} \mathcal{L}(\beta, \phi)$ inside the q-norm $\|.\|_q$ defined above, while we rely on standard index and matrix notation for all other expressions involving those partial derivatives, e.g. $\partial_{\phi\phi'\phi_g} \mathcal{L}(\beta, \phi)$ is a $\dim \phi \times \dim \phi$ matrix for every $g = 1, \ldots, \dim \phi$. Occasionally, e.g. in Assumption B.1(vi), we use the Euclidean norm for $\dim \phi$-vectors, and the spectral norm for $\dim \phi \times \dim \phi$-matrices, denoted by $\|.\|$, and defined as $\|.\|_q$ with $q = 2$. Moreover, we employ the matrix infinity norm $\|A\|_\infty = \max_i \sum_j |A_{ij}|$, and the matrix maximum norm $\|A\|_{\max} = \max_{ij} |A_{ij}|$ to characterize the properties of the inverse of the expected Hessian of the incidental parameters in Appendix D.

For $r \geq 0$, we define the sets $\mathcal{B}(r, \beta^0) = \{\beta : \|\beta - \beta^0\| \leq r\}$, and $\mathcal{B}_q(r, \phi^0) = \{\phi : \|\phi - \phi^0\|_q \leq r\}$, which are closed balls of radius r around the true parameter values β^0 and ϕ^0, respectively.

Appendix B. Asymptotic expansions

In this section, we derive asymptotic expansions for the score of the profile objective function, $\mathcal{L}(\beta, \widehat{\phi}(\beta))$, and for the fixed effects estimators of the parameters and APEs, $\widehat{\beta}$ and $\widehat{\delta}$. We do not employ the panel structure of the model, nor the particular form of the objective function given in Section 4. Instead, we consider the estimation of an unspecified model based on a sample of size NT and a generic objective function $\mathcal{L}(\beta, \phi)$, which depends on the parameter of interest β and the incidental parameter ϕ. The estimators $\widehat{\phi}(\beta)$ and $\widehat{\beta}$ are defined in (2.7) and (2.8). The proof of all the results in this Section are given in the supplementary material (see Appendix E).

We make the following high-level assumptions. These assumptions might appear somewhat abstract, but will be justified by more primitive conditions in the context of panel models.

Assumption B.1 (*Regularity Conditions for Asymptotic Expansion of* $\widehat{\beta}$). Let $q > 4$ and $0 \leq \epsilon < 1/8 - 1/(2q)$. Let $r_\beta = r_{\beta,NT} > 0$, $r_\phi = r_{\phi,NT} > 0$, with $r_\beta = o\left[(NT)^{-1/(2q)-\epsilon}\right]$ and $r_\phi = o\left[(NT)^{-\epsilon}\right]$. We assume that

(i) $\frac{\dim \phi}{\sqrt{NT}} \to a$, $0 < a < \infty$.
(ii) $(\beta, \phi) \mapsto \mathcal{L}(\beta, \phi)$ is four times continuously differentiable in $\mathcal{B}(r_\beta, \beta^0) \times \mathcal{B}_q(r_\phi, \phi^0)$, wpa1.

(iii) $\displaystyle\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\left\|\widehat{\phi}(\beta)-\phi^0\right\|_q = o_P(r_\phi).$

(iv) $\overline{\mathcal{H}} > 0$, and $\left\|\overline{\mathcal{H}}^{-1}\right\|_q = \mathcal{O}_P(1).$

(v) For the q-norm defined in Appendix A,

$\|\mathcal{S}\|_q = \mathcal{O}_P\left((NT)^{-1/4+1/(2q)}\right), \qquad \|\partial_\beta\mathcal{L}\| = \mathcal{O}_P(1),$

$\left\|\widetilde{\mathcal{H}}\right\|_q = o_P(1),$

$\left\|\partial_{\beta\phi'}\mathcal{L}\right\|_q = \mathcal{O}_P\left((NT)^{1/(2q)}\right),$

$\left\|\partial_{\beta\beta'}\mathcal{L}\right\| = \mathcal{O}_P(\sqrt{NT}), \qquad \left\|\partial_{\beta\phi\phi}\mathcal{L}\right\|_q = \mathcal{O}_P((NT)^\epsilon),$

$\left\|\partial_{\phi\phi\phi}\mathcal{L}\right\|_q = \mathcal{O}_P\left((NT)^\epsilon\right),$

and

$\displaystyle\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi^0)}\left\|\partial_{\beta\beta\beta}\mathcal{L}(\beta,\phi)\right\| = \mathcal{O}_P\left(\sqrt{NT}\right),$

$\displaystyle\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi^0)}\left\|\partial_{\beta\beta\phi}\mathcal{L}(\beta,\phi)\right\|_q = \mathcal{O}_P\left((NT)^{1/(2q)}\right),$

$\displaystyle\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi^0)}\left\|\partial_{\beta\beta\phi\phi}\mathcal{L}(\beta,\phi)\right\|_q = \mathcal{O}_P\left((NT)^\epsilon\right),$

$\displaystyle\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi^0)}\left\|\partial_{\beta\phi\phi\phi}\mathcal{L}(\beta,\phi)\right\|_q = \mathcal{O}_P\left((NT)^\epsilon\right),$

$\displaystyle\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\sup_{\phi\in\mathcal{B}_q(r_\phi,\phi^0)}\left\|\partial_{\phi\phi\phi\phi}\mathcal{L}(\beta,\phi)\right\|_q = \mathcal{O}_P\left((NT)^\epsilon\right).$

(vi) For the spectral norm $\|.\| = \|.\|_2$,

$\left\|\widetilde{\mathcal{H}}\right\| = o_P\left((NT)^{-1/8}\right), \qquad \left\|\partial_{\beta\beta'}\widetilde{\mathcal{L}}\right\| = o_P(\sqrt{NT}),$

$\left\|\partial_{\beta\phi\phi}\widetilde{\mathcal{L}}\right\| = o_P\left((NT)^{-1/8}\right),$

$\left\|\partial_{\beta\phi'}\widetilde{\mathcal{L}}\right\| = \mathcal{O}_P(1),$

$\left\|\sum_{g,h=1}^{\dim\phi}\partial_{\phi\phi_g\phi_h}\widetilde{\mathcal{L}}[\overline{\mathcal{H}}^{-1}\mathcal{S}]_g[\overline{\mathcal{H}}^{-1}\mathcal{S}]_h\right\| = o_P\left((NT)^{-1/4}\right).$

Let $\partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta))$ be the score of the profile objective function.[20] The following theorem is the main result of this Appendix.

Theorem B.1 (*Asymptotic Expansions of $\widehat{\phi}(\beta)$ and $\partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta))$*). *Let Assumption B.1 hold. Then*

$$\widehat{\phi}(\beta)-\phi^0 = \mathcal{H}^{-1}\mathcal{S} + \mathcal{H}^{-1}[\partial_{\phi\beta'}\mathcal{L}](\beta-\beta^0)$$
$$+ \frac{1}{2}\mathcal{H}^{-1}\sum_{g=1}^{\dim\phi}[\partial_{\phi\phi'\phi_g}\mathcal{L}]\mathcal{H}^{-1}\mathcal{S}[\mathcal{H}^{-1}\mathcal{S}]_g + R^\phi(\beta),$$

and

$$\partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta)) = U - \overline{W}\sqrt{NT}(\beta-\beta^0) + R(\beta),$$

where $U = U^{(0)} + U^{(1)}$, and

$$\overline{W} = -\frac{1}{\sqrt{NT}}\left(\partial_{\beta\beta'}\overline{\mathcal{L}} + [\partial_{\beta\phi'}\overline{\mathcal{L}}]\,\overline{\mathcal{H}}^{-1}\,[\partial_{\phi\beta'}\overline{\mathcal{L}}]\right),$$

$$U^{(0)} = \partial_\beta\mathcal{L} + [\partial_{\beta\phi'}\overline{\mathcal{L}}]\,\overline{\mathcal{H}}^{-1}\mathcal{S},$$

$$U^{(1)} = [\partial_{\beta\phi'}\widetilde{\mathcal{L}}]\overline{\mathcal{H}}^{-1}\mathcal{S} - [\partial_{\beta\phi'}\overline{\mathcal{L}}]\,\overline{\mathcal{H}}^{-1}\,\widetilde{\mathcal{H}}\,\overline{\mathcal{H}}^{-1}\mathcal{S}$$
$$+ \frac{1}{2}\sum_{g=1}^{\dim\phi}\left(\partial_{\beta\phi'\phi_g}\overline{\mathcal{L}} + [\partial_{\beta\phi'}\overline{\mathcal{L}}]\,\overline{\mathcal{H}}^{-1}[\partial_{\phi\phi'\phi_g}\overline{\mathcal{L}}]\right)[\overline{\mathcal{H}}^{-1}\mathcal{S}]_g\,\overline{\mathcal{H}}^{-1}\mathcal{S}.$$

---

[20] Note that $\frac{d}{d\beta}\mathcal{L}(\beta,\widehat{\phi}(\beta)) = \partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta))$ by the envelope theorem.

The remainder terms of the expansions satisfy

$$\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\frac{(NT)^{1/2-1/(2q)}\left\|R^\phi(\beta)\right\|_q}{1+\sqrt{NT}\|\beta-\beta^0\|} = o_P(1),$$

$$\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\frac{\|R(\beta)\|}{1+\sqrt{NT}\|\beta-\beta^0\|} = o_P(1).$$

Remark 8. The result for $\widehat{\phi}(\beta)-\phi^0$ does not rely on Assumption B.1(vi). Without this assumption we can also show that

$$\partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta)) = \partial_\beta\mathcal{L} + \left[\partial_{\beta\beta'}\mathcal{L} + (\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}(\partial_{\phi'\beta}\mathcal{L})\right](\beta-\beta^0)$$
$$+ (\partial_{\beta\phi'}\mathcal{L})\mathcal{H}^{-1}\mathcal{S} + \frac{1}{2}\sum_g\left(\partial_{\beta\phi'\phi_g}\mathcal{L}\right.$$
$$+ \left.[\partial_{\beta\phi'}\mathcal{L}]\,\mathcal{H}^{-1}[\partial_{\phi\phi'\phi_g}\mathcal{L}]\right)[\mathcal{H}^{-1}\mathcal{S}]_g\mathcal{H}^{-1}\mathcal{S} + R_1(\beta),$$

with $R_1(\beta)$ satisfying the same bound as $R(\beta)$. Thus, the spectral norm bounds in Assumption B.1(vi) for $\dim\phi$-vectors, matrices and tensors are only used after separating expectations from deviations of expectations for certain partial derivatives. Otherwise, the derivation of the bounds is purely based on the q-norm for $\dim\phi$-vectors, matrices and tensors.

The proofs are given in Section S.3 of the supplementary material. Theorem B.1 characterizes asymptotic expansions for the incidental parameter estimator and the score of the profile objective function in the incidental parameter score $\mathcal{S}$ up to quadratic order. The theorem provides bounds on the remainder terms $R^\phi(\beta)$ and $R(\beta)$, which make the expansions applicable to estimators of β that take values within a shrinking r_β-neighborhood of β^0 wpa1. Given such an r_β-consistent estimator $\widehat{\beta}$ that solves the first order condition $\partial_\beta\mathcal{L}(\beta,\widehat{\phi}(\beta)) = 0$, we can use the expansion of the profile objective score to obtain an asymptotic expansion for $\widehat{\beta}$. This gives rise to the following corollary of Theorem B.1. Let $\overline{W}_\infty := \lim_{N,T\to\infty}\overline{W}$.

Corollary B.2 (*Asymptotic Expansion of $\widehat{\beta}$*). *Let Assumption B.1 be satisfied. In addition, let $U = \mathcal{O}_P(1)$, let $\overline{W}_\infty$ exist with $\overline{W}_\infty > 0$, and let $\|\widehat{\beta}-\beta^0\| = o_P(r_\beta)$. Then*

$$\sqrt{NT}(\widehat{\beta}-\beta^0) = \overline{W}_\infty^{-1}U + o_P(1).$$

The following theorem states that for strictly concave objective functions no separate consistency proof is required for $\widehat{\phi}(\beta)$ and for $\widehat{\beta}$.

Theorem B.3 (*Consistency under Concavity*). *Let Assumption B.1(i), (ii), (iv), (v) and (vi) hold, and let $(\beta,\phi) \mapsto \mathcal{L}(\beta,\phi)$ be strictly concave over $(\beta,\phi) \in \mathbb{R}^{\dim\beta+\dim\phi}$, wpa1. Assume furthermore that $(NT)^{-1/4+1/(2q)} = o_P(r_\phi)$ and $(NT)^{1/(2q)}r_\beta = o_P(r_\phi)$. Then,*

$$\sup_{\beta\in\mathcal{B}(r_\beta,\beta^0)}\left\|\widehat{\phi}(\beta)-\phi^0\right\|_q = o_P(r_\phi),$$

i.e. Assumption B.1(iii) is satisfied. If, in addition, $\overline{W}_\infty$ exists with $\overline{W}_\infty > 0$, then $\|\widehat{\beta}-\beta^0\| = \mathcal{O}_P\left((NT)^{-1/4}\right)$.

In the application of Theorem B.1 to panel models, we focus on estimators with strictly concave objective functions. By Theorem B.3, we only need to check Assumption B.1(i), (ii), (iv), (v) and (vi), as well as $U = \mathcal{O}_P(1)$ and $\overline{W}_\infty > 0$, when we apply Corollary B.2 to derive the limiting distribution of $\widehat{\beta}$. We give the proofs of Corollary B.2 and Theorem B.3 in Section S.3.

Expansion for average effects

We invoke the following high-level assumption, which is verified under more primitive conditions for panel data models in the next section.

Assumption B.2 (*Regularity Conditions for Asymptotic Expansion of* $\widehat{\delta}$). Let q, ϵ, r_β and r_ϕ be defined as in Assumption B.1. We assume that

(i) $(\beta, \phi) \mapsto \Delta(\beta, \phi)$ is three times continuously differentiable in $\mathcal{B}(r_\beta, \beta^0) \times \mathcal{B}_q(r_\phi, \phi^0)$, wpa1.
(ii) $\|\partial_\beta \Delta\| = \mathcal{O}_P(1)$, $\|\partial_\phi \Delta\|_q = \mathcal{O}_P\left((NT)^{1/(2q)-1/2}\right)$, $\|\partial_{\phi\phi} \Delta\|_q = \mathcal{O}_P((NT)^{\epsilon-1/2})$, and

$$\sup_{\beta \in \mathcal{B}(r_\beta, \beta^0)} \sup_{\phi \in \mathcal{B}_q(r_\phi, \phi^0)} \|\partial_{\beta\beta} \Delta(\beta, \phi)\| = \mathcal{O}_P(1),$$

$$\sup_{\beta \in \mathcal{B}(r_\beta, \beta^0)} \sup_{\phi \in \mathcal{B}_q(r_\phi, \phi^0)} \|\partial_{\beta\phi'} \Delta(\beta, \phi)\|_q = \mathcal{O}_P\left((NT)^{1/(2q)-1/2}\right),$$

$$\sup_{\beta \in \mathcal{B}(r_\beta, \beta^0)} \sup_{\phi \in \mathcal{B}_q(r_\phi, \phi^0)} \|\partial_{\phi\phi} \Delta(\beta, \phi)\|_q = \mathcal{O}_P\left((NT)^{\epsilon-1/2}\right).$$

(iii) $\|\partial_\beta \widetilde{\Delta}\| = o_P(1)$, $\|\partial_\phi \widetilde{\Delta}\| = \mathcal{O}_P\left((NT)^{-1/2}\right)$, and $\|\partial_{\phi\phi} \widetilde{\Delta}\| = o_P\left((NT)^{-5/8}\right)$.

The following result gives the asymptotic expansion for the estimator, $\widehat{\delta} = \Delta(\beta, \widehat{\phi}(\beta))$, wrt $\delta = \Delta(\beta^0, \phi^0)$.

Theorem B.4 (*Asymptotic Expansion of* $\widehat{\delta}$). Let Assumptions B.1 and B.2 hold and let $\|\widehat{\beta} - \beta^0\| = \mathcal{O}_P\left((NT)^{-1/2}\right) = o_P(r_\beta)$. Then

$$\widehat{\delta} - \delta = \left[\partial_{\beta'} \overline{\Delta} + (\partial_{\phi'} \overline{\Delta}) \overline{\mathcal{H}}^{-1} (\partial_{\phi\beta'} \overline{\mathcal{L}})\right] (\widehat{\beta} - \beta^0) + U_\Delta^{(0)}$$
$$+ U_\Delta^{(1)} + o_P\left(1/\sqrt{NT}\right),$$

where

$$U_\Delta^{(0)} = (\partial_{\phi'} \overline{\Delta}) \overline{\mathcal{H}}^{-1} \mathscr{S},$$
$$U_\Delta^{(1)} = (\partial_{\phi'} \widetilde{\Delta}) \overline{\mathcal{H}}^{-1} \mathscr{S} - (\partial_{\phi'} \overline{\Delta}) \overline{\mathcal{H}}^{-1} \widetilde{\mathcal{H}} \overline{\mathcal{H}}^{-1} \mathscr{S}$$
$$+ \frac{1}{2} \mathscr{S}' \overline{\mathcal{H}}^{-1} \left[\partial_{\phi\phi'} \overline{\Delta} + \sum_{g=1}^{\dim \phi} \left[\partial_{\phi\phi'\phi_g} \overline{\mathcal{L}}\right] \left[\overline{\mathcal{H}}^{-1} (\partial_\phi \overline{\Delta})\right]_g\right] \overline{\mathcal{H}}^{-1} \mathscr{S}.$$

Remark 9. The expansion of the profile score $\partial_{\beta_k} \mathcal{L}(\beta, \widehat{\phi}(\beta))$ in Theorem B.1 is a special case of the expansion in Theorem B.4, for $\Delta(\beta, \phi) = \frac{1}{\sqrt{NT}} \partial_{\beta_k} \mathcal{L}(\beta, \phi)$. Assumption B.2 also exactly matches with the corresponding subset of Assumption B.1.

Appendix C. Proofs of Section 4

C.1. Application of general expansion to panel estimators

We now apply the general expansion of Appendix B to the panel fixed effects estimators considered in the main text. For the objective function specified in (2.1) and (4.1), the incidental parameter score evaluated at the true parameter value is

$$\mathscr{S} = \left(\begin{bmatrix}\frac{1}{\sqrt{NT}} \sum_{t=1}^T \partial_\pi \ell_{it}\end{bmatrix}_{i=1,\dots,N} \atop \begin{bmatrix}\frac{1}{\sqrt{NT}} \sum_{i=1}^N \partial_\pi \ell_{it}\end{bmatrix}_{t=1,\dots,T}\right).$$

The penalty term in the objective function does not contribute to $\mathscr{S}$, because at the true parameter value $v'\phi^0 = 0$. The

corresponding expected incidental parameter Hessian $\overline{\mathcal{H}}$ is given in (4.2). Appendix D discusses the structure of $\overline{\mathcal{H}}$ and $\overline{\mathcal{H}}^{-1}$ in more detail. Define

$$\Lambda_{it} := -\frac{1}{\sqrt{NT}} \sum_{j=1}^N \sum_{\tau=1}^T \left(\overline{\mathcal{H}}^{-1}_{(\alpha\alpha)ij} + \overline{\mathcal{H}}^{-1}_{(\gamma\alpha)tj}\right.$$
$$\left.+ \overline{\mathcal{H}}^{-1}_{(\alpha\gamma)i\tau} + \overline{\mathcal{H}}^{-1}_{(\gamma\gamma)t\tau}\right) \partial_\pi \ell_{j\tau}, \tag{C.1}$$

and the operator $D_\beta \Delta_{it} := \partial_\beta \Delta_{it} - \partial_\pi \Delta_{it} \Xi_{it}$, which are similar to Ξ_{it} and $D_\beta \ell_{it}$ in Eq. (4.3).

The following theorem shows that Assumptions 4.1 and 4.2 for the panel model are sufficient for Assumptions B.1 and B.2 for the general expansion, and particularizes the terms of the expansion to the panel estimators. The proof is given in the supplementary material (see Appendix E).

Theorem C.1. *Consider an estimator with objective function given by* (2.1) *and* (4.1). *Let Assumption 4.1 be satisfied and suppose that the limit* $\overline{W}_\infty$ *defined in Theorem 4.1 exists and is positive definite. Let* $q = 8$, $\epsilon = 1/(16 + 2\nu)$, $r_{\beta,NT} = \log(NT)(NT)^{-1/8}$ *and* $r_{\phi,NT} = (NT)^{-1/16}$. *Then,*

(i) *Assumption B.1 holds and* $\|\widehat{\beta} - \beta^0\| = \mathcal{O}_P((NT)^{-1/4})$.
(ii) *The approximate Hessian and the terms of the score defined in Theorem B.1 can be written as*

$$\overline{W} = -\frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \mathbb{E}_\phi \left(\partial_{\beta\beta'} \ell_{it} - \partial_{\pi^2} \ell_{it} \Xi_{it} \Xi'_{it}\right),$$

$$U^{(0)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T D_\beta \ell_{it},$$

$$U^{(1)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^N \sum_{t=1}^T \left\{-\Lambda_{it} \left[D_{\beta\pi} \ell_{it} - \mathbb{E}_\phi(D_{\beta\pi} \ell_{it})\right]\right.$$
$$\left.+ \frac{1}{2} \Lambda_{it}^2 \mathbb{E}_\phi(D_{\beta\pi^2} \ell_{it})\right\}.$$

(iii) *In addition, let Assumption 4.2 hold. Then, Assumption B.2 is satisfied for the partial effects defined in* (2.2). *By Theorem B.4,*

$$\sqrt{NT}\left(\widehat{\delta} - \delta\right) = V_\Delta^{(0)} + V_\Delta^{(1)} + o_P(1),$$

where

$$V_\Delta^{(0)} = \left[\frac{1}{NT} \sum_{i,t} \mathbb{E}_\phi(D_\beta \Delta_{it})\right]' \overline{W}_\infty^{-1} U^{(0)}$$
$$- \frac{1}{\sqrt{NT}} \sum_{i,t} \mathbb{E}_\phi(\Psi_{it}) \partial_\pi \ell_{it},$$

$$V_\Delta^{(1)} = \left[\frac{1}{NT} \sum_{i,t} \mathbb{E}_\phi(D_\beta \Delta_{it})\right]' \overline{W}_\infty^{-1} U^{(1)}$$
$$+ \frac{1}{\sqrt{NT}} \sum_{i,t} \Lambda_{it} \left[\Psi_{it} \partial_{\pi^2} \ell_{it} - \mathbb{E}_\phi(\Psi_{it}) \mathbb{E}_\phi(\partial_{\pi^2} \ell_{it})\right]$$
$$+ \frac{1}{2\sqrt{NT}} \sum_{i,t} \Lambda_{it}^2 \left[\mathbb{E}_\phi(\partial_{\pi^2} \Delta_{it}) - \mathbb{E}_\phi(\partial_{\pi^3} \ell_{it}) \mathbb{E}_\phi(\Psi_{it})\right].$$

C.2. Proofs of Theorems 4.1 and 4.2

Proof of Theorem 4.1. # First, we want to show that $U^{(0)} \to_d \mathcal{N}(0, \overline{W}_\infty)$. In our likelihood setting, $\mathbb{E}_\phi \partial_\beta \mathcal{L} = 0$, $\mathbb{E}_\phi \mathscr{S} = 0$, and, by the Bartlett identities, $\mathbb{E}_\phi(\partial_\beta \mathcal{L} \partial_{\beta'} \mathcal{L}) = -\frac{1}{\sqrt{NT}} \partial_{\beta\beta'} \overline{\mathcal{L}}$, $\mathbb{E}_\phi(\partial_\beta \mathcal{L} \mathscr{S}') = -\frac{1}{\sqrt{NT}} \partial_{\beta\phi'} \overline{\mathcal{L}}$ and $\mathbb{E}_\phi(\mathscr{S}\mathscr{S}') = \frac{1}{\sqrt{NT}}\left(\overline{\mathcal{H}} - \frac{b}{\sqrt{NT}} vv'\right)$.

Furthermore, $\mathcal{S}'v = 0$ and $\partial_{\beta\phi'}\overline{\mathcal{L}}v = 0$. Then, by definition of $\overline{W} = -\frac{1}{\sqrt{NT}}\left(\partial_{\beta\beta'}\overline{\mathcal{L}} + [\partial_{\beta\phi'}\overline{\mathcal{L}}]\,\overline{\mathcal{H}}^{-1}\,[\partial_{\phi\beta'}\overline{\mathcal{L}}]\right)$ and $U^{(0)} = \partial_{\beta}\mathcal{L} + [\partial_{\beta\phi'}\overline{\mathcal{L}}]\,\overline{\mathcal{H}}^{-1}\mathcal{S}$,

$$\mathbb{E}_{\phi}\left(U^{(0)}\right) = 0, \qquad \mathrm{Var}\left(U^{(0)}\right) = \overline{W},$$

which implies that $\lim_{N,T\to\infty}\mathrm{Var}\left(U^{(0)}\right) = \lim_{N,T\to\infty}\overline{W} = \overline{W}_{\infty}$. Moreover, part (ii) of Theorem C.1 yields

$$U^{(0)} = \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}D_{\beta}\ell_{it},$$

where $D_{\beta}\ell_{it} = \partial_{\beta}\ell_{it} - \partial_{\pi}\ell_{it}\,\Xi_{it}$ is a martingale difference sequence for each i and independent across i, conditional on ϕ. Thus, by Lemma S.3 and the Cramer–Wold device we conclude that

$$U^{(0)} \to_d \mathcal{N}\left[0, \lim_{N,T\to\infty}\mathrm{Var}\left(U^{(0)}\right)\right] \sim \mathcal{N}(0, \overline{W}_{\infty}).$$

Next, we show that $U^{(1)} \to_P \kappa\overline{B}_{\infty} + \kappa^{-1}\overline{D}_{\infty}$. Part (ii) of Theorem C.1 gives $U^{(1)} = U^{(1a)} + U^{(1b)}$, with

$$U^{(1a)} = -\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda_{it}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right],$$

$$U^{(1b)} = \frac{1}{2\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\Lambda_{it}^{2}\,\mathbb{E}_{\phi}(D_{\beta\pi^2}\ell_{it}).$$

Plugging-in the definition of Λ_{it}, we decompose $U^{(1a)} = U^{(1a,1)} + U^{(1a,2)} + U^{(1a,3)} + U^{(1a,4)}$, where

$$U^{(1a,1)} = \frac{1}{NT}\sum_{i,j}\overline{\mathcal{H}}_{(\alpha\alpha)ij}^{-1}\left(\sum_{\tau}\partial_{\pi}\ell_{j\tau}\right)\sum_{t}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right],$$

$$U^{(1a,2)} = \frac{1}{NT}\sum_{j,t}\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}\left(\sum_{\tau}\partial_{\pi}\ell_{j\tau}\right)\sum_{i}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right],$$

$$U^{(1a,3)} = \frac{1}{NT}\sum_{i,\tau}\overline{\mathcal{H}}_{(\alpha\gamma)i\tau}^{-1}\left(\sum_{j}\partial_{\pi}\ell_{j\tau}\right)\sum_{t}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right],$$

$$U^{(1a,4)} = \frac{1}{NT}\sum_{t,\tau}\overline{\mathcal{H}}_{(\gamma\gamma)t\tau}^{-1}\left(\sum_{j}\partial_{\pi}\ell_{j\tau}\right)\sum_{i}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right].$$

By the Cauchy–Schwarz inequality applied to the sum over t in $U^{(1a,2)}$,

$$\left(U^{(1a,2)}\right)^2 \le \frac{1}{(NT)^2}\left[\sum_{t}\left(\sum_{j,\tau}\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}\partial_{\pi}\ell_{j\tau}\right)^2\right]$$
$$\times \left[\sum_{t}\left(\sum_{i}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right]\right)^2\right].$$

By Lemma D.1, $\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1} = \mathcal{O}_P(1/\sqrt{NT})$, uniformly over t, j. Using that both $\sqrt{NT}\,\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}\partial_{\pi}\ell_{j\tau}$ and $D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})$ are mean zero, independent across i and Lemma S.2 in the supplementary material across t, we obtain

$$\mathbb{E}_{\phi}\left(\frac{1}{\sqrt{NT}}\sum_{j,\tau}[\sqrt{NT}\,\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}]\partial_{\pi}\ell_{j\tau}\right)^2 = \mathcal{O}_P(1),$$

$$\mathbb{E}_{\phi}\left(\frac{1}{\sqrt{N}}\sum_{i}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right]\right)^2 = \mathcal{O}_P(1),$$

uniformly over t. Thus, $\sum_{t}\left(\sum_{j,\tau}\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}\partial_{\pi}\ell_{j\tau}\right)^2 = \mathcal{O}_P(T)$ and $\sum_{t}\left(\sum_{i}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right]\right)^2 = \mathcal{O}_P(NT)$. We conclude that

$$\left(U^{(1a,2)}\right)^2 = \frac{1}{(NT)^2}\mathcal{O}_P(T)\mathcal{O}_P(NT) = \mathcal{O}_P(1/N) = o_P(1),$$

and therefore that $U^{(1a,2)} = o_P(1)$. Analogously one can show that $U^{(1a,3)} = o_P(1)$.

By Lemma D.1, $\overline{\mathcal{H}}_{(\alpha\alpha)}^{-1} = -\mathrm{diag}\left[\left(\frac{1}{\sqrt{NT}}\sum_{t=1}^{T}\mathbb{E}_{\phi}(\partial_{\pi^2}\ell_{it})\right)^{-1}\right] + \mathcal{O}_P(1/\sqrt{NT})$. Analogously to the proof of $U^{(1a,2)} = o_P(1)$, one can show that the $\mathcal{O}_P(1/\sqrt{NT})$ part of $\overline{\mathcal{H}}_{(\alpha\alpha)}^{-1}$ has an asymptotically negligible contribution to $U^{(1a,1)}$. Thus,

$$U^{(1a,1)} = -\frac{1}{\sqrt{NT}}\sum_{i}\underbrace{\frac{\left(\sum_{\tau}\partial_{\pi}\ell_{i\tau}\right)\sum_{t}\left[D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})\right]}{\sum_{t}\mathbb{E}_{\phi}(\partial_{\pi^2}\ell_{it})}}_{=:U_i^{(1a,1)}}$$
$$+ o_P(1).$$

Our assumptions guarantee that $\mathbb{E}_{\phi}\left[\left(U_i^{(1a,1)}\right)^2\right] = \mathcal{O}_P(1)$, uniformly over i. Note that both the denominator and the numerator of $U_i^{(1a,1)}$ are of order T. For the denominator this is obvious because of the sum over T. For the numerator there are two sums over T, but both $\partial_{\pi}\ell_{i\tau}$ and $D_{\beta\pi}\ell_{it} - \mathbb{E}_{\phi}(D_{\beta\pi}\ell_{it})$ are mean zero weakly correlated processes, so that their sums are of order $\sqrt{T}$. By the WLLN over i (remember that we have cross-sectional independence, conditional on ϕ, and we assume finite moments), $N^{-1}\sum_i U_i^{(1a,1)} = N^{-1}\sum_i \mathbb{E}_{\phi}U_i^{(1a,1)} + o_P(1)$, and therefore

$$U^{(1a,1)} = -\sqrt{\frac{N}{T}}\frac{1}{N}\sum_{i=1}^{N}\underbrace{\frac{\sum_{t=1}^{T}\sum_{\tau=t}^{T}\mathbb{E}_{\phi}\left(\partial_{\pi}\ell_{it}D_{\beta\pi}\ell_{i\tau}\right)}{\sum_{t=1}^{T}\mathbb{E}_{\phi}\left(\partial_{\pi^2}\ell_{it}\right)}}_{=:\sqrt{\frac{N}{T}}\overline{B}^{(1)}} + o_P(1).$$

Here, we use that $\mathbb{E}_{\phi}\left(\partial_{\pi}\ell_{it}D_{\beta\pi}\ell_{i\tau}\right) = 0$ for $t > \tau$. Analogously,

$$U^{(1a,4)} = -\sqrt{\frac{T}{N}}\frac{1}{T}\sum_{t=1}^{T}\underbrace{\frac{\sum_{i=1}^{N}\mathbb{E}_{\phi}\left(\partial_{\pi}\ell_{it}D_{\beta\pi}\ell_{it}\right)}{\sum_{i=1}^{N}\mathbb{E}_{\phi}\left(\partial_{\pi^2}\ell_{it}\right)}}_{=:\sqrt{\frac{T}{N}}\overline{D}^{(1)}} + o_P(1).$$

We conclude that $U^{(1a)} = \kappa\overline{B}^{(1)} + \kappa^{-1}\overline{D}^{(1)} + o_P(1)$.

Next, we analyze $U^{(1b)}$. We decompose $\Lambda_{it} = \Lambda_{it}^{(1)} + \Lambda_{it}^{(2)} + \Lambda_{it}^{(3)} + \Lambda_{it}^{(4)}$, where

$$\Lambda_{it}^{(1)} = -\frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\overline{\mathcal{H}}_{(\alpha\alpha)ij}^{-1}\sum_{\tau=1}^{T}\partial_{\pi}\ell_{j\tau},$$

$$\Lambda_{it}^{(2)} = -\frac{1}{\sqrt{NT}}\sum_{j=1}^{N}\overline{\mathcal{H}}_{(\gamma\alpha)tj}^{-1}\sum_{\tau=1}^{T}\partial_{\pi}\ell_{j\tau},$$

$$\Lambda_{it}^{(3)} = -\frac{1}{\sqrt{NT}}\sum_{\tau=1}^{T}\overline{\mathcal{H}}_{(\alpha\gamma)i\tau}^{-1}\sum_{\tau=1}^{T}\partial_{\pi}\ell_{j\tau},$$

$$\Lambda_{it}^{(4)} = -\frac{1}{\sqrt{NT}} \sum_{\tau=1}^{T} \overline{\mathcal{H}}_{(\gamma\gamma)t\tau}^{-1} \sum_{\tau=1}^{T} \partial_\pi \ell_{j\tau}.$$

This decomposition of Λ_{it} induces the following decomposition of $U^{(1b)}$

$$U^{(1b)} = \sum_{p,q=1}^{4} U^{(1b,p,q)},$$

$$U^{(1b,p,q)} = \frac{1}{2\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \Lambda_{it}^{(p)} \Lambda_{it}^{(q)} \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it}).$$

Due to the symmetry $U^{(1b,p,q)} = U^{(1b,q,p)}$, this decomposition has 10 distinct terms. Start with $U^{(1b,1,2)}$ noting that

$$U^{(1b,1,2)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} U_i^{(1b,1,2)},$$

$$U_i^{(1b,1,2)} = \frac{1}{2T} \sum_{t=1}^{T} \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it}) \frac{1}{N^2} \sum_{j_1,j_2=1}^{N} \left[NT\overline{\mathcal{H}}_{(\alpha\alpha)ij_1}^{-1} \overline{\mathcal{H}}_{(\gamma\alpha)tj_2}^{-1} \right]$$
$$\times \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_1\tau} \right) \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_2\tau} \right).$$

By $\mathbb{E}_\phi(\partial_\pi \ell_{it}) = 0$, $\mathbb{E}_\phi(\partial_\pi \ell_{it} \partial_\pi \ell_{j\tau}) = 0$ for $(i,t) \neq (j,\tau)$, and the properties of the inverse expected Hessian from Lemma D.1, $\mathbb{E}_\phi\left[U_i^{(1b,1,2)}\right] = \mathcal{O}_P(1/N)$, uniformly over i, $\mathbb{E}_\phi\left[\left(U_i^{(1b,1,2)}\right)^2\right] = \mathcal{O}_P(1)$, uniformly over i, and $\mathbb{E}_\phi\left[U_i^{(1b,1,2)} U_j^{(1b,1,2)}\right] = \mathcal{O}_P(1/N)$, uniformly over $i \neq j$. This implies that $\mathbb{E}_\phi U^{(1b,1,2)} = \mathcal{O}_P(1/N)$ and $\mathbb{E}_\phi\left[\left(U^{(1b,1,2)} - \mathbb{E}_\phi U^{(1b,1,2)}\right)^2\right] = \mathcal{O}_P(1/\sqrt{N})$, and therefore $U^{(1b,1,2)} = o_P(1)$. By similar arguments one obtains $U^{(1b,p,q)} = o_P(1)$ for all combinations of $p,q = 1,2,3,4$, except for $p = q = 1$ and $p = q = 4$.

For $p = q = 1$,

$$U^{(1b,1,1)} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} U_i^{(1b,1,1)},$$

$$U_i^{(1b,1,1)} = \frac{1}{2T} \sum_{t=1}^{T} \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it}) \frac{1}{N^2} \sum_{j_1,j_2=1}^{N} \left[NT\overline{\mathcal{H}}_{(\alpha\alpha)ij_1}^{-1} \overline{\mathcal{H}}_{(\alpha\alpha)ij_2}^{-1} \right]$$
$$\times \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_1\tau} \right) \left(\frac{1}{\sqrt{T}} \sum_{\tau=1}^{T} \partial_\pi \ell_{j_2\tau} \right).$$

Analogous to the result for $U^{(1b,1,2)}$, $\mathbb{E}_\phi\left[\left(U^{(1b,1,1)} - \mathbb{E}_\phi U^{(1b,1,1)}\right)^2\right] = \mathcal{O}_P(1/\sqrt{N})$, and therefore $U^{(1b,1,1)} = \mathbb{E}_\phi U^{(1b,1,1)} + o(1)$. Furthermore,

$$\mathbb{E}_\phi U^{(1b,1,1)} = \frac{1}{2\sqrt{NT}} \sum_{i=1}^{N} \frac{\sum_{t=1}^{T} \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it}) \sum_{\tau=1}^{T} \mathbb{E}_\phi\left[(\partial_\pi \ell_{i\tau})^2\right]}{\left[\sum_{t=1}^{T} \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right) \right]^2} + o(1)$$

$$= -\sqrt{\frac{N}{T}} \frac{1}{2N} \sum_{i=1}^{N} \frac{\sum_{t=1}^{T} \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it})}{\sum_{t=1}^{T} \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)} + o(1).$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad}_{=:\sqrt{\frac{N}{T}}\overline{B}^{(2)}}$$

Analogously,

$$U^{(1b,4,4)} = \mathbb{E}_\phi U^{(1b,4,4)} + o_P(1)$$

$$= -\sqrt{\frac{T}{N}} \frac{1}{2T} \sum_{t=1}^{T} \frac{\sum_{i=1}^{N} \mathbb{E}_\phi(D_{\beta\pi^2}\ell_{it})}{\sum_{i=1}^{N} \mathbb{E}_\phi\left(\partial_{\pi^2}\ell_{it}\right)} + o(1).$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad}_{=:\sqrt{\frac{T}{N}}\overline{D}^{(2)}}$$

We have thus shown that $U^{(1b)} = \kappa\overline{B}^{(2)} + \kappa^{-1}\overline{D}^{(2)} + o_P(1)$. Since $\overline{B}_\infty = \lim_{N,T\to\infty}[\overline{B}^{(1)} + \overline{B}^{(2)}]$ and $\overline{D}_\infty = \lim_{N,T\to\infty}[\overline{D}^{(1)} + \overline{D}^{(2)}]$ we thus conclude $U^{(1)} = \kappa\overline{B}_\infty + \kappa^{-1}\overline{D}_\infty + o_P(1)$.

We have shown $U^{(0)} \to_d \mathcal{N}(0, \overline{W}_\infty)$, and $U^{(1)} \to_P \kappa\overline{B}_\infty + \kappa^{-1}\overline{D}_\infty$. Then, part (ii) of Theorem C.1 yields $\sqrt{NT}(\widehat{\beta} - \beta^0) \to_d \overline{W}_\infty^{-1} \mathcal{N}(\kappa\overline{B}_\infty + \kappa^{-1}\overline{D}_\infty, \overline{W}_\infty)$. ∎

Proof of Theorem 4.2. We consider the case of scalar Δ_{it} to simplify the notation. Decompose

$$r_{NT}(\widehat{\delta} - \delta_{NT}^0 - \overline{B}_\infty^\delta/T - \overline{D}_\infty^\delta/N) = r_{NT}(\delta - \delta_{NT}^0)$$
$$+ \frac{r_{NT}}{\sqrt{NT}}\sqrt{NT}(\widehat{\delta} - \delta - \overline{B}_\infty^\delta/T - \overline{D}_\infty^\delta/N).$$

Part (1): Limit of $\sqrt{NT}(\widehat{\delta} - \delta - \overline{B}_\infty^\delta/T - \overline{D}_\infty^\delta/N)$. An argument analogous to the proof of Theorem 4.1 using Theorem C.1(iii) yields

$$\sqrt{NT}(\widehat{\delta} - \delta) \to_d \mathcal{N}\left(\kappa\overline{B}_\infty^\delta + \kappa^{-1}\overline{D}_\infty^\delta, \overline{V}_\infty^{\delta(1)}\right),$$

where $\overline{V}_\infty^{\delta(1)} = \text{plim}_{N,T\to\infty}\left\{(NT)^{-1}\sum_{i,t}\mathbb{E}_\phi[\Gamma_{it}^2]\right\}$, for the expressions of $\overline{B}_\infty^\delta$, $\overline{D}_\infty^\delta$, and Γ_{it} given in the statement of the theorem. Then, by Mann–Wald theorem

$$\sqrt{NT}(\widehat{\delta} - \delta - \overline{B}_\infty^\delta/T - \overline{D}_\infty^\delta/N) \to_d \mathcal{N}\left(0, \overline{V}_\infty^{\delta(1)}\right).$$

Part (2): Limit of $r_{NT}(\delta - \delta_{NT}^0)$. Here we show that $r_{NT}(\delta - \delta_{NT}^0) \to_d \mathcal{N}(0, \overline{V}_\infty^{\delta(2)})$ for the convergence rate r_{NT} given in Remark 4, and characterize the asymptotic variance $\overline{V}_\infty^{\delta(2)}$. We determine r_{NT} through $\mathbb{E}[(\delta - \delta_{NT}^0)^2] = \mathcal{O}(r_{NT}^{-2})$ and $r_{NT}^{-2} = \mathcal{O}(\mathbb{E}[(\delta - \delta_{NT}^0)^2])$, where

$$\mathbb{E}[(\delta - \delta_{NT}^0)^2] = \mathbb{E}\left[\left(\frac{1}{NT}\sum_{i,t}\widetilde{\Delta}_{it}\right)^2\right]$$
$$= \frac{1}{N^2T^2}\sum_{i,j,t,s}\mathbb{E}\left[\widetilde{\Delta}_{it}\widetilde{\Delta}_{js}\right], \qquad (C.2)$$

for $\widetilde{\Delta}_{it} = \Delta_{it} - \mathbb{E}(\Delta_{it})$. Then, we characterize $\overline{V}_\infty^{\delta(2)}$ as $\overline{V}_\infty^{\delta(2)} = \text{plim}_{N,T\to\infty}\{r_{NT}^2\mathbb{E}[(\delta - \delta_{NT}^0)^2]\}$, because $\mathbb{E}[\delta - \delta_{NT}^0] = 0$. The order of $\mathbb{E}[(\delta - \delta_{NT}^0)^2]$ is equal to the number of terms of the sums in Eq. (C.2) that are non zero, which is determined by the sample properties of $\{(X_{it}, \alpha_i, \gamma_t) : 1 \leq i \leq N, 1 \leq t \leq T\}$. Under Assumption 4.2(i), if $\{\alpha_i\}_N$ and $\{\gamma_t\}_T$ are independent sequences, and α_i and γ_t are independent for all i, t, then $\mathbb{E}[\widetilde{\Delta}_{it}\widetilde{\Delta}_{js}] = \mathbb{E}[\widetilde{\Delta}_{it}]\mathbb{E}[\widetilde{\Delta}_{js}] = 0$ if $i \neq j$ and $t \neq s$, so that

$$\mathbb{E}[(\delta - \delta_{NT}^0)^2] = \frac{1}{N^2T^2}\left\{\sum_{i,t,s}\mathbb{E}\left[\widetilde{\Delta}_{it}\widetilde{\Delta}_{is}\right] + \sum_{i,j,t}\mathbb{E}\left[\widetilde{\Delta}_{it}\widetilde{\Delta}_{jt}\right]\right.$$
$$\left. - \sum_{i,t}\mathbb{E}\left[\widetilde{\Delta}_{it}^2\right]\right\} = \mathcal{O}\left(\frac{N+T-1}{NT}\right),$$

because $\mathbb{E}[\widetilde{\Delta}_{it}\,\widetilde{\Delta}_{is}] \leq \mathbb{E}[\mathbb{E}_\phi(\widetilde{\Delta}_{it}^2)]^{1/2}\mathbb{E}[\mathbb{E}_\phi(\widetilde{\Delta}_{is}^2)]^{1/2} < C$ by the Cauchy–Schwarz inequality and Assumption 4.2(ii). We conclude that $r_{NT} = \sqrt{NT/(N+T-1)}$ and

$$\overline{V}^{\delta(2)} = \plim_{N,T\to\infty} \left\{ \frac{r_{NT}^2}{N^2T^2} \left(\sum_{i,t,s} \mathbb{E}\left[\widetilde{\Delta}_{it}\,\widetilde{\Delta}_{is}\right] + \sum_{i\neq j,t} \mathbb{E}\left[\widetilde{\Delta}_{it}\,\widetilde{\Delta}_{jt}\right] \right) \right\}.$$

Note that $r_{NT} \to \infty$ and $r_{NT} = \mathcal{O}(\sqrt{NT})$.

\# Part (3): limit of $r_{NT}(\widehat{\delta} - \delta_{NT}^0 - T^{-1}\overline{B}_\infty^\delta - N^{-1}\overline{D}_\infty^\delta)$. The conclusion of the Theorem follows because $(\delta - \delta_{NT}^0)$ and $(\widehat{\delta} - \delta - T^{-1}\overline{B}_\infty^\delta - N^{-1}\overline{D}_\infty^\delta)$ are asymptotically independent and $\overline{V}_\infty^\delta = \overline{V}^{\delta(2)} + \overline{V}^{\delta(1)} \lim_{N,T\to\infty}(r_{NT}/\sqrt{NT})^2$. ∎

Appendix D. Properties of the inverse expected incidental parameter hessian

The expected incidental parameter Hessian evaluated at the true parameter values is

$$\overline{\mathcal{H}} = \mathbb{E}_\phi[-\partial_{\phi\phi'}\mathcal{L}] = \begin{pmatrix} \overline{\mathcal{H}}_{(\alpha\alpha)}^* & \overline{\mathcal{H}}_{(\alpha\gamma)}^* \\ [\overline{\mathcal{H}}_{(\alpha\gamma)}^*]' & \overline{\mathcal{H}}_{(\gamma\gamma)}^* \end{pmatrix} + \frac{b}{\sqrt{NT}}\, vv',$$

where $v = v_{NT} = (1_N', -1_T')'$, $\overline{\mathcal{H}}_{(\alpha\alpha)}^* = \mathrm{diag}(\frac{1}{\sqrt{NT}} \sum_t \mathbb{E}_\phi[-\partial_{\pi^2}\ell_{it}])$, $\overline{\mathcal{H}}_{(\alpha\gamma)it}^* = \frac{1}{\sqrt{NT}}\mathbb{E}_\phi[-\partial_{\pi^2}\ell_{it}]$, and $\overline{\mathcal{H}}_{(\gamma\gamma)}^* = \mathrm{diag}(\frac{1}{\sqrt{NT}} \sum_i \mathbb{E}_\phi[-\partial_{\pi^2}\ell_{it}])$.

In panel models with only individual effects, it is straightforward to determine the order of magnitude of $\overline{\mathcal{H}}^{-1}$ in Assumption B.1(iv), because $\overline{\mathcal{H}}$ contains only the diagonal matrix $\overline{\mathcal{H}}_{(\alpha\alpha)}^*$. In our case, $\overline{\mathcal{H}}$ is no longer diagonal, but it has a special structure. The diagonal terms are of order 1, whereas the off-diagonal terms are of order $(NT)^{-1/2}$. Moreover, $\left\|\overline{\mathcal{H}} - \mathrm{diag}(\overline{\mathcal{H}}_{(\alpha\alpha)}^*, \overline{\mathcal{H}}_{(\gamma\gamma)}^*)\right\|_{\max} = \mathcal{O}_P((NT)^{-1/2})$. These observations, however, are not sufficient to establish the order of $\overline{\mathcal{H}}^{-1}$ because the number of non-zero off-diagonal terms is of much larger order than the number of diagonal terms; compare $\mathcal{O}(NT)$ to $\mathcal{O}(N+T)$. Note also that the expected Hessian without penalty term $\overline{\mathcal{H}}^*$ has the same structure as $\overline{\mathcal{H}}$ itself, but is not even invertible, i.e. the observation on the relative size of diagonal vs. off-diagonal terms is certainly not sufficient to make statements about the structure of $\overline{\mathcal{H}}^{-1}$. The result of the following lemma is therefore not obvious. It shows that the diagonal terms of $\overline{\mathcal{H}}$ also dominate in determining the order of $\overline{\mathcal{H}}^{-1}$.

Lemma D.1. *Under* Assumption 4.1,

$$\left\|\overline{\mathcal{H}}^{-1} - \mathrm{diag}\left(\overline{\mathcal{H}}_{(\alpha\alpha)}^*, \overline{\mathcal{H}}_{(\gamma\gamma)}^*\right)^{-1}\right\|_{\max} = \mathcal{O}_P\left((NT)^{-1/2}\right).$$

The proof of Lemma D.1 is provided in the supplementary material (see Appendix E). The lemma result establishes that $\overline{\mathcal{H}}^{-1}$ can be uniformly approximated by a diagonal matrix, which is given by the inverse of the diagonal terms of $\overline{\mathcal{H}}$ without the penalty. The diagonal elements of $\mathrm{diag}(\overline{\mathcal{H}}_{(\alpha\alpha)}^*, \overline{\mathcal{H}}_{(\gamma\gamma)}^*)^{-1}$ are of order 1, i.e. the order of the difference established by the lemma is relatively small.

Note that the choice of penalty in the objective function is important to obtain Lemma D.1. Different penalties, corresponding to other normalizations (e.g. a penalty proportional to α_1^2, corresponding to the normalization $\alpha_1 = 0$), would fail to deliver Lemma D.1. However, these alternative choices do not affect the estimators $\widehat{\beta}$ and $\widehat{\delta}$, i.e. which normalization is used to compute $\widehat{\beta}$ and $\widehat{\delta}$ in practice is irrelevant (up to numerical precision errors).

Appendix E. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.12.014.

References

Aghion, P., Bloom, N., Blundell, R., Griffith, R., Howitt, P., 2005. Competition and innovation: an inverted-U relationship. Quart. J. Econ. 120 (2), 701–728.

Alvarez, J., Arellano, M., 2003. The time series and cross-section asymptotics of dynamic panel data estimators. Econometrica 71 (4), 1121–1159.

Arellano, M., Bonhomme, S., 2009. Robust priors in nonlinear panel data models. Econometrica 77 (2), 489–536.

Arellano, M., Hahn, J., 2007. Advances in economics and econometrics theory and applications. In: Ninth World Congress, Econometric Society Monographs, Cambridge University Press, chapter "Understanding Bias in Nonlinear Panel Models: Some Recent Developments", vol. 3. pp. 381–409.

Bai, J., 2009. Panel data models with interactive fixed effects. Econometrica 77 (4), 1229–1279.

Carro, J., 2007. Estimating dynamic panel data discrete choice models with fixed effects. J. Econometrics 140 (2), 503–528.

Chamberlain, G., 2010. Binary response models for panel data: Identification and information. Econometrica 78 (1), 159–168.

Charbonneau, K., 2012. Multiple fixed effects in nonlinear panel data models (unpublished manuscript).

Charbonneau, K., 2014. Multiple Fixed Effects in Binary Response Panel Data Models. Bank of Canada Working Paper 2014-17.

Chen, M., Fernandez-Val, I., Weidner, M., 2014. Nonlinear Panel Models with Interactive Effects. ArXiv e-prints.

Chernozhukov, V., Fernández-Val, I., Hahn, J., Newey, W., 2013. Average and quantile effects in nonseparable panel models. Econometrica 81 (2), 535–580.

Cruz-González, M., Fernández-Val, I., Weidner, M., 2015. probitfe and logitfe: Bias corrections for fixed effects estimators of probit and logit panel models with individual and time effects (unpublished manuscript).

Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effect models. Rev. Econom. Stud. 82 (3), 991–1030.

Enea, M., 2012. Speedglm: Fitting Generalized Linear and Generalized Linear Models to large data sets. R package version 0.1.

Fernández-Val, I., 2009. Fixed effects estimation of structural parameters and marginal effects in panel probit models. J. Econometrics 150, 71–85.

Fernández-Val, I., Lee, J., 2013. Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. Quant. Econ. 4 (3), 453–481.

Fernández-Val, I., Vella, F., 2011. Bias corrections for two-step fixed effects panel data estimators. J. Econometrics 163 (2), 144–162.

Fernández-Val, I., Weidner, M., 2015a. Individual and Time Effects in Nonlinear Panel Models with Large N, T. Cemmap Working Paper, Centre for Microdata Methods and Practice.

Fernández-Val, I., Weidner, M., 2015b. Supplement to 'Individual and time effects in nonlinear panel models with large N,T' (unpublished manuscript).

Galvao, A.F., Kato, K., 2014. Estimation and inference for linear panel data models under misspecification when both N and T are large. J. Bus. Econom. Statist. 32 (2), 285–309.

Greene, W., 2004. The behavior of the fixed effects estimator in nonlinear models. Econom. J. 7 (1), 98–119.

Hahn, J., Kuersteiner, G., 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., Kuersteiner, G., 2007. Bandwidth Choice for Bias Estimators in Dynamic Nonlinear Panel Models. Working Paper.

Hahn, J., Kuersteiner, G., 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. Econometric Theory 27 (06), 1152–1191.

Hahn, J., Moon, H., 2006. Reducing bias of MLE in a dynamic panel model. Econometric Theory 22 (03), 499–512.

Hahn, J., Newey, W., 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72 (4), 1295–1319.

Heckman, J., 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: Structural Analysis of Discrete Data with Econometric Applications. pp. 179–195.

Honoré, B.E., Tamer, E., 2006. Bounds on parameters in panel dynamic discrete choice models. Econometrica 611–629.

Hu, L., 2002. Estimation of a censored dynamic panel data model. Econometrica 70 (6), 2499–2517.

Kato, K., Galvao, A., Montes-Rojas, G., 2012. Asymptotics for panel quantile regression models with individual effects. J. Econometrics 170 (1), 76–91.

Kristensen, D., Salanié, B., 2013. Higher-Order Properties of Approximate Estimators. Cemmap Working Paper, Centre for Microdata Methods and Practice.

Lancaster, T., 2000. The incidental parameter problem since 1948. J. Econometrics 95 (2), 391–413.

Lancaster, T., 2002. Orthogonal parameters and panel data. Rev. Econom. Stud. 69 (3), 647–666.

Moon, H., Weidner, M., 2015a. Dynamic linear panel regression models with interactive fixed effects. Econometric Theory (forthcoming).

Moon, H.R., Weidner, M., 2015b. Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica 83 (4), 1543–1579.

Neyman, J., Scott, E., 1948. Consistent estimates based on partially consistent observations. Econometrica 16 (1), 1–32.

Okui, R., 2013. Asymptotically unbiased estimation of autocovariances and autocorrelations with panel data in the presence of individual and time effects. J. Time Ser. Econ. 1–53.

Olsen, R., 1978. Note on the uniqueness of the maximum likelihood estimator for the tobit model. Econometrica 1211–1215.

Pesaran, M.H., 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica 74 (4), 967–1012.

Phillips, P.C. B., Moon, H., 1999. Linear regression limit theory for nonstationary panel data. Econometrica 67 (5), 1057–1111.

Pratt, J.W., 1981. Concavity of the log likelihood. J. Amer. Statist. Assoc. 76 (373), 103–106.

Woutersen, T., 2002. Robustness against incidental parameters (unpublished manuscript).