# INTEGRATED LIKELIHOOD BASED INFERENCE FOR NONLINEAR PANEL DATA MODELS WITH UNOBSERVED EFFECTS

MARTIN SCHUMANN, THOMAS A. SEVERINI, AND GAUTAM TRIPATHI

Abstract. Panel data models with fixed effects are widely used by economists and other social scientists to capture the effects of unobserved individual heterogeneity. In this paper, we propose a new integrated likelihood based approach for estimating panel data models when the unobserved individual effects enter the model nonlinearly. Unlike existing integrated likelihoods in the literature, the one we propose is closer to a genuine likelihood. Although the statistical theory for the proposed estimator is developed in an asymptotic setting where the number of individuals and the number of time periods both approach infinity, results from a simulation study suggest that our methodology can work very well even in moderately sized panels of short duration in both static and dynamic models.

## 1. introduction

A panel dataset, also known as a longitudinal dataset, consists of observations on individuals recorded over a period of time. This data structure is rich enough to allow economists and other social scientists to estimate and test models of economic and social outcomes that contain as explanatory variables not only individual characteristics observed by the researchers, but also attributes unobserved to the researchers that vary across individuals but not across time. Since these unobserved individual characteristics may be correlated with some or all of the observed individual characteristics, simply ignoring them can lead to severely biased statistical inference. Time-invariant (at least in the short-run) characteristics such as ability, productivity, or latent cultural preferences, which vary only at the individual level and cannot be seen by researchers analyzing the data, are referred to as "unobserved individual heterogeneity," "unobserved individual effects," or simply as "fixed effects," in order to distinguish them from variables such as gender and ethnicity, which are also time-invariant but typically observed by researchers at the individual level. In this paper, we use the term "fixed effects."

To justify the time-invariance of fixed effects, panel datasets in microeconometric applications are often characterized as being "short," i.e., having the number of individuals ($n$) much larger than the number of time periods ($T$). It is therefore not surprising that, in many papers, the asymptotic theory of inference for microeconometric panel data models is developed

Source: il-rev-2j.tex. Compiled: 14th August, 2020 at 10:23pm.

*Key words and phrases.* Fixed effects, Integrated likelihood, Nonlinear models, Panel data.

under the assumption that $n \to \infty$ and $T$ is held fixed. In this setting, it is well known how to estimate and test models with fixed effects without making distributional assumptions when the fixed effects enter the model linearly and additively; cf., e.g., Chamberlain (1982, 1984), Hsiao (1986), Baltagi (1995), Arellano (2003b), and Wooldridge (2010). By contrast, parametric distributional assumptions are usually needed to estimate models where the fixed effects enter nonlinearly. Even then, there is only a limited class of nonlinear models that can be satisfactorily studied in a unified manner due to the well known "incidental parameters problem" (Lancaster, 2000). For instance, while it is known how to consistently estimate a panel logit with fixed effects (Andersen, 1970; Chamberlain, 1980), the same approach does not work for estimating probit models, a closely related specification (Magnac, 2004; Chamberlain, 2010).

The objective of this paper is to develop a unified methodology, based on a new integrated likelihood (IL) approach, for estimating nonlinear panel data models with fixed effects in a parametric likelihood framework. The IL we propose can be regarded as extending the IL of Lancaster (2002) and Arellano and Bonhomme (2009). The statistical theory justifying the proposed methodology is developed in a setting where both $n$ and $T$ are allowed to grow such that $n$ grows faster than $T$; i.e., we also focus on short panels, although the validity of our approximations is guaranteed only when $n, T \to \infty$.

The paper is organized as follows. Section 2 outlines the model. Section 3 compares our work with Lancaster (2002) and Arellano and Bonhomme (2009). Sections 4–7 describe our estimation approach and show that our IL possesses some desirable properties. Section 8 develops the asymptotic theory for the proposed estimator. Section 9 illustrates how our approach works in some familiar settings, and Section 10 investigates its small sample properties for logit, probit, and AR(1) designs. Section 11 concludes. Implementation details, additional simulation results, technical assumptions and their justification, and all figures and proofs, are in Appendices A–J, available as supplementary material for this paper.

## 2. THE MODEL

In this section, we specify our likelihood based model, the basic notation, and the sampling and identification assumptions maintained throughout the paper. Technical assumptions used to derive the results in this paper are in Appendix C.

Let $Y_{it}$ denote outcomes and $X_{it}$ a vector of explanatory variables for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. Hereafter, $n, T \geq 2$ and "vector" means a column vector. The random variables $Y_{it}$ and $X_{it}$ are observed, with $i$ indexing the individual and $t$ the time. The fixed effect $\alpha_{i0}$ is an unobserved random variable whose distribution is unknown.

**Assumption 2.1** (Fixed effects)**.** *For each $i$, the unobserved random variable $\alpha_{i0}$ is continuously distributed with support $(\mathfrak{a}, \mathfrak{b})$, where $\mathfrak{a}, \mathfrak{b} \in \mathbb{R} \cup \{-\infty, +\infty\}$, $\mathfrak{a} < \mathfrak{b}$, and $\mathfrak{a}, \mathfrak{b}$ are known.*

Excluding Section 8, where we require $(\mathfrak{a}, \mathfrak{b})$ to be a bounded interval in order to show consistency of our estimator (cf. Assumption C.5(vi) and the discussion following it), we allow $(\mathfrak{a}, \mathfrak{b}) = \mathbb{R}$ in the remainder of the paper. In particular, $(\mathfrak{a}, \mathfrak{b}) = \mathbb{R}$ is maintained for all examples in Section 9 and all simulation designs in Section 10. Henceforth, let $\mathcal{Y}_{iT} := (Y_{i1}, \ldots, Y_{iT})$ denote the time-series of outcomes, and $\mathcal{X}_{iT} := (X_{i1}, \ldots, X_{iT})$ the time-series of explanatory variables, corresponding to the $i$th individual for the duration of the panel.

The distribution of $(\mathcal{X}_{iT}, \alpha_{i0})$ is unknown, which allows for arbitrary correlation between the fixed effects and the explanatory variables. Given $(\mathcal{X}_{iT}, \alpha_{i0})$, the time-series $\mathcal{Y}_{iT}$ is drawn from the conditional density $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$, which is known up to a parameter $\theta_0 \in \text{int}(\Theta)$, where $\Theta$ is a known subset of $\mathbb{R}^{\dim(\theta_0)}$ with nonempty interior. It is assumed that $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$ is a density with respect to an appropriate dominating measure (Lebesgue, counting, or a mixture of both), which does not depend on $(\mathcal{X}_{iT}, \alpha_{i0}, \theta_0)$; the dominating measure is, therefore, not explicitly specified.

**Assumption 2.2** (Identification)**.** *$\theta_0$ is identified, i.e., uniquely defined.*

In particular, as specified in Assumption C.5(vii), $\theta_0$ is identified as the well-separated global maximum of the limit (as $n, T \to \infty$) of the expected "target" loglikelihood (defined in Section 3) for the sample. The presence of $\alpha_{i0}$ implies that parameters associated with observed time-invariant explanatory variables are, in general, not identifiable. For instance, suppose that a vector of observed time-invariant explanatory variables $Z_i$ enters the model density as $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},b_i(Z_i,\delta_0,\alpha_{i0});\theta_0}$, where the time-invariant function $b_i$ takes values in $(\mathfrak{a}, \mathfrak{b})$ and is known up to a finite dimensional parameter $\delta_0$, e.g., $b_i(Z_i, \delta_0, \alpha_{i0}) = Z_i'\delta_0 + \alpha_{i0}$ provided $Z_i'\delta_0 + \alpha_{i0} \in (\mathfrak{a}, \mathfrak{b})$. Then, $\delta_0$ is not identified because, for each $T$, the density $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},b_i(Z_i,\delta_0,\alpha_{i0});\theta_0}$ is observationally equivalent to the density $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\tilde{\alpha}_{i0};\theta_0}$, where $\tilde{\alpha}_{i0} := b_i(Z_i, \delta_0, \alpha_{i0}) \in (\mathfrak{a}, \mathfrak{b})$ is another fixed effect. Consequently, Assumption 2.2 rules out time-invariant explanatory variables in $\mathcal{X}_{iT}$ that have a time-invariant relationship with $\alpha_{i0}$.

**Assumption 2.3** (Sampling)**.** (i) *For each $T$, $(\mathcal{Y}_{1T}, \mathcal{X}_{1T}, \alpha_{10}), \ldots, (\mathcal{Y}_{nT}, \mathcal{X}_{nT}, \alpha_{n0})$ are independently and identically distributed (i.i.d.)*; (ii) *For each $i$, conditional on $\mathcal{X}_{iT}, \alpha_{i0}$, the outcomes $Y_{i1}, \ldots, Y_{iT}$ are independent*; (iii) *For each $i$, conditional on $\alpha_{i0}$, the process $(Y_{it}, X_{it})_{t\in\mathbb{N}}$ is strictly stationary.*

(i), which stipulates that observations across $i$ are i.i.d., is typical in microeconometric applications and is maintained throughout the paper. (ii) imposes conditional independence within each $i$, hereafter referred to as "time-independence," which rules out lagged outcomes as explanatory variables, and is maintained everywhere except in Section 9.2. Although some of the assumptions in Appendix C.2–C.4 are justified under time-independence, it is, in principle, not necessary for the methodology developed in this paper to work. To illustrate this,

we apply our approach to the linear dynamic panel data model in Section 9.2 without imposing time-independence. (iii) rules out the presence of time-varying parameters or time-trends in the model for the outcomes, i.e., there are no time-varying parameters or time-trends in $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$. The stationarity assumption is also maintained, e.g., in Hahn and Kuersteiner (2002, 2011), Hahn and Newey (2004), Arellano and Bonhomme (2009), and Dhaene and Jochmans (2015). Allowing for time-varying parameters in our approach is, in principle, possible. For instance, Mikailov (2017) has demonstrated that our approach goes through in the Neyman-Scott model (Example B.1) in the presence of both time and fixed effects. However, doing so will significantly increase the technical complexity of the proofs in the $n, T \to \infty$ setting, cf., e.g., Fernández-Val and Weidner (2016), without affecting the raison d'être of the methodology we propose. Therefore, (iii) is maintained throughout the paper.

Since $\alpha_{i0}$ is an unobserved random variable, we can talk about the likelihood of a potential realization. Specifically, if $\theta \in \Theta$ and $\alpha_i \in (\mathfrak{a}, \mathfrak{b})$ denotes a possible value taken by $\alpha_{i0}$, then we define the likelihood of $(\theta, \alpha_i)$ for the $i$th individual to be $L_{iT}(\theta, \alpha_i) := f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_i;\theta}(\mathcal{Y}_{iT})$. The average loglikelihood for the $i$th individual is denoted by $\ell_{iT}(\theta, \alpha_i) := T^{-1} \log L_{iT}(\theta, \alpha_i)$. We will refer to $\theta$ as the parameter of interest, and call $\alpha_i$ an individual specific nuisance parameter. The loglikelihood function $(\theta, \alpha_i) \mapsto \ell_{iT}(\theta, \alpha_i)$ is assumed to be sufficiently well-behaved so that derivatives with respect to $(\theta, \alpha_i)$, as many as needed in the paper, can be interchanged with integrals respect to the density $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_i;\theta}$, and the mixed partial derivatives are equal.

The score of $\ell_{iT}(\theta, \alpha_i)$ with respect to $\theta$ is the (column) vector $\ell_{iT\theta}(\theta, \alpha_i) := \nabla_\theta \ell_{iT}(\theta, \alpha_i)$, where $\nabla_\theta := (\partial_\theta)'$ is the gradient and "$'$" the transpose operator. Similarly, as $\alpha_i$ is a scalar, $\ell_{iT\alpha}(\theta, \alpha_i) := \nabla_\alpha \ell_{iT}(\theta, \alpha_i) = \partial_\alpha \ell_{iT}(\theta, \alpha_i)$ denotes the score with respect to $\alpha_i$. We use $f_{ab} := \partial_b \circ \nabla_a f$ to denote mixed partial derivatives of second order. Consequently, $\ell_{iT\theta\theta}(\theta, \alpha_i)$ is a square matrix, $\ell_{iT\theta\alpha}(\theta, \alpha_i)$ is a column vector, $\ell_{iT\alpha\theta}(\theta, \alpha_i)$ is a row vector (with $\ell'_{iT\alpha\theta} = \ell_{iT\theta\alpha}$), and $\ell_{iT\alpha\alpha}(\theta, \alpha_i)$ is a scalar. Third (and higher) order partial derivatives, when at most one derivative with respect to $\theta$ is taken, are defined analogously. For instance, $\ell_{iT\theta\alpha\alpha}(\theta, \alpha_i) := \partial_\alpha^2 \circ \nabla_\theta \ell_{iT}(\theta, \alpha_i)$ is a column vector, $\ell_{iT\alpha\alpha\theta}(\theta, \alpha_i) := \partial_\theta \circ \partial_\alpha \circ \nabla_\alpha \ell_{iT}(\theta, \alpha_i)$ and $\ell_{iT\alpha\theta\alpha}(\theta, \alpha_i) := \partial_\alpha \circ \partial_\theta \circ \nabla_\alpha \ell_{iT}(\theta, \alpha_i)$ are row vectors, and $\ell_{iT\alpha\alpha\alpha\alpha}(\theta, \alpha_i) := \partial_\alpha^3 \circ \nabla_\alpha \ell_{iT}(\theta, \alpha_i)$ is a scalar.

Given $(\check{\theta}, \check{\alpha}) \in \Theta \times (\mathfrak{a}, \mathfrak{b})$, let $\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \check{\theta}, \check{\alpha}] := \int_{\text{supp}(\mathcal{Y}_{iT})} \ell_{iT\alpha}(\theta, \alpha_i) f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\check{\alpha};\check{\theta}}$, where $\text{supp}(\mathcal{Y}_{iT})$ denotes the support of $\mathcal{Y}_{iT}$, and integration is with respect to the (unspecified) dominating measure for $f$. The integral can be calculated analytically or numerically, depending on the functional form of $f$. Integration with respect to $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$ is denoted by $\mathbb{E}_0$, e.g., $\mathbb{E}_0 \ell_{iT\alpha}(\theta, \alpha_i) := \mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \theta_0, \alpha_{i0}]$. Similarly, $\text{var}_0\, \ell_{iT\alpha}(\theta, \alpha_i) := \mathbb{E}_0[\ell_{iT\alpha}(\theta, \alpha_i) - \mathbb{E}_0 \ell_{iT\alpha}(\theta, \alpha_i)]^2$. We usually omit the arguments in functionals of $\mathbb{E}_0 \ell_{iT}(\theta, \alpha_i)$ when evaluated at $(\theta_0, \alpha_{i0})$. E.g., we write $\mathbb{E}_0 \ell_{iT\alpha}^2 := \mathbb{E}_0 \ell_{iT\alpha}^2(\theta_0, \alpha_{i0})$, $\mathbb{E}_0 \ell_{iT\theta\alpha} := \mathbb{E}_0 \ell_{iT\theta\alpha}(\theta_0, \alpha_{i0})$, $\mathbb{E}_0 \ell_{iT\alpha\alpha} \ell_{iT\alpha} := \mathbb{E}_0 \ell_{iT\alpha\alpha}(\theta_0, \alpha_{i0}) \ell_{iT\alpha}(\theta_0, \alpha_{i0})$, etc.. Since $\mathbb{E}_0$ is a conditional expectation and $\text{var}_0$ a conditional variance (both conditional on $\mathcal{X}_{iT}, \alpha_{i0}$), equalities and inequalities involving them hold w.p.1.

This is the sense in which subsequent statements and assumptions regarding $\mathbb{E}_0$ and $\text{var}_0$ should be interpreted even when the "w.p.1" qualifier is missing. To avoid the proliferation of "w.p.1" qualifiers each time $\mathbb{E}_0$ or $\text{var}_0$ is mentioned, we do not state them explicitly hereafter.

Henceforth, to allow the number of time periods to grow simultaneously with the number of individuals, let $(T_n)$ be a sequence of positive integers such that $T_n \to \infty$ as $n \to \infty$. When there is no danger of confusion, we do not indicate the dependence of estimators on $n$ and $T$. If $A_{iT}$ is an array, then the statement $A_{iT} = O_{\text{p}}(1)$ is understood to hold coordinatewise.

## 3. PREVIOUS WORK AND OUR CONTRIBUTION

Since the distribution of $\mathcal{Y}_{iT}|\mathcal{X}_{iT}, \alpha_i$ is known up to $(\theta, \alpha_i)$, it is natural to estimate $\theta$ by maximum likelihood while treating $\alpha_i$ as a nuisance parameter. The fixed effects maximum likelihood estimator (MLE) of $\theta$ is given by

$$\tilde{\theta} := \underset{\theta \in \Theta}{\text{argmax}} \max_{\alpha_1, \ldots, \alpha_n \in (\mathfrak{a}, \mathfrak{b})} n^{-1} \sum_{i=1}^{n} \ell_{iT}(\theta, \alpha_i). \tag{3.1}$$

Unfortunately, since the number of nuisance parameters grows with the number of individuals, the MLE of $\theta$ may not be consistent when $n \to \infty$ and $T$ is fixed. Example B.1 in the supplementary material, due to Neyman and Scott (1948), illustrates this phenomenon beautifully.

Since each individual contributes a single individual specific nuisance parameter, the MLE in (3.1) can be written as $\tilde{\theta} = \text{argmax}_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \ell_{iT}^{\text{p}}(\theta)$, where $\ell_{iT}^{\text{p}}(\theta) := \ell_{iT}(\theta, \hat{\alpha}_{iT}(\theta))$ is the average profile loglikelihood of $\theta$ for individual $i$ after the nuisance parameters have been concentrated out, and

$$\hat{\alpha}_{iT}(\theta) := \underset{u \in (\mathfrak{a}, \mathfrak{b})}{\text{argmax}}\, \ell_{iT}(\theta, u), \qquad \theta \in \Theta, \tag{3.2}$$

is the MLE of $\alpha_i$ for a given $\theta$. Henceforth, for future reference, we let

$$\alpha_{iT}^{*}(\theta) := \underset{u \in (\mathfrak{a}, \mathfrak{b})}{\text{argmax}}\, \mathbb{E}_0 \ell_{iT}(\theta, u), \qquad \theta \in \Theta, \tag{3.3}$$

and refer to it as the "population level MLE" of $\alpha_i$ for a given $\theta$. Following Pace and Salvan (2006, Section 3.2), we refer to $\ell_{iT}(\theta, \alpha_{iT}^{*}(\theta))$ as the "target" loglikelihood of $\theta$ for the $i$th individual. The (infeasible) target loglikelihood satisfies all of the Bartlett identities because it is a genuine loglikelihood.

The inconsistency of $\tilde{\theta}$ (when $n \to \infty$, $T$ is fixed, and the profile likelihood is not a conditional likelihood free of nuisance parameters as in Example 9.2), is due to the fact that the individual specific nuisance parameters are poorly estimated when $T$ is held fixed. Indeed, from its definition it is clear that $\hat{\alpha}_{iT}(\theta_0)$ does not depend on $n$. Therefore, when $\hat{\alpha}_{iT}(\theta_0) \neq \alpha_{i0}$, $\hat{\alpha}_{iT}(\theta_0)$ will not converge to $\alpha_{i0}$ when $n \to \infty$ and $T$ is fixed. This is the sense in which the individual specific parameters are poorly estimated when $n \to \infty$ but $T$ is fixed. This implies that the profile likelihood scores for each individual do not have zero mean when evaluated at

the true parameter values (McCullagh and Tibshirani, 1990, Remark 2, p. 329). In fact, it is shown in Appendix B that

$$\mathbb{E}_0 \nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-1}), \tag{3.4}$$

and that $\tilde{\theta}$ is inconsistent, as $n \to \infty$ and $T$-fixed, where $\nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta) = \ell_{iT\theta}(\theta, \hat{\alpha}_{iT}(\theta))$ is the profile likelihood score of $\theta$ for the $i$th individual.

Since the score function of a genuine loglikelihood has zero mean, (3.4) reveals that the bias of $\nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta_0)$ is of the order $1/T$ for each $i$. Hence, allowing $T$ to grow (along with $n$) may enable $\tilde{\theta}$ to consistently estimate its true value $\theta_0$ as both $n, T \to \infty$. However, as is clear from the Neyman-Scott model in Example B.2, this alone may not be sufficient to ensure that $\tilde{\theta}$ is asymptotically unbiased in the sense that the limiting distribution of $\sqrt{nT_n}(\tilde{\theta} - \theta_0)$ is correctly centered at the origin.

Research to solve this problem, namely, to construct an estimator whose asymptotic distribution is correctly centered and whose asymptotic variance equals that of the fixed effects MLE, has generated a large literature. E.g., Lancaster (2002) has suggested an IL approach based on orthogonalizing the parameter of interest and the individual specific nuisance parameter using the "information orthogonalizing transformation (IOT);" cf. Cox and Reid (1987) and Severini (2000, Section 3.6.4) on how the IOT is obtained. Lancaster's approach can be restrictive because the IOT may not exist when $\theta$ is a vector. Important applications where this can happen include the AR(1) model with covariates (Lancaster, 2002, Section 3.2), and autoregressive models of order greater than one (this was conjectured by Lancaster, 2002, p. 663, who did not provide a proof; a proof can be found in Dhaene and Jochmans, 2016, p. 1208).

It is thus desirable to obtain estimators of $\theta$ that do not require the IOT, so that they are applicable in general situations where $\theta$ is a vector. One approach is to employ the jackknife or analytical bias corrections. Cf., e.g., Hahn and Kuersteiner (2002), Woutersen (2002), Arellano (2003a), Hahn and Newey (2004), Arellano and Hahn (2007), Carro (2007), Bester and Hansen (2009), Fernández-Val (2009), Hahn and Kuersteiner (2011), and Dhaene and Jochmans (2015). Alternatively, Arellano and Bonhomme (2009), henceforth AB, propose an IL approach that does not require $\theta$ and $\alpha_i$ to be orthogonal. Their estimator is $\hat{\theta}_{\mathrm{AB}} := \operatorname{argmax}_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \bar{\ell}^{\mathrm{AB}}_{iT}(\theta)$, where $\bar{\ell}^{\mathrm{AB}}_{iT}(\theta) := T^{-1} \log \int L_{iT}(\theta, \alpha) \hat{w}_i(\theta, \alpha)\, d\alpha$ is the log-IL of AB, and $\hat{w}_i$ is an individual specific data-dependent weight-function. The weights $\hat{w}_1, \ldots, \hat{w}_n$ are chosen such that the bias of the IL score for each $i$, when each $\hat{w}_i$ is replaced by its population counterparts and all parameters are evaluated at the truth, is of the order $1/T^2$ as $T \to \infty$. For each $i$, the population scores in AB's approach can therefore be regarded as being "first-order unbiased" as compared to the profile likelihood scores, whose bias is only of the order $1/T$ (cf. (3.4)). Under the condition that $\lim_{n\to\infty} n/T_n \in (0, \infty)$, $\sqrt{nT_n}(\hat{\theta}_{\mathrm{AB}} - \theta_0)$ is asymptotically normal with mean zero and variance equal to that of the fixed effects MLE. Recent works

similar to AB include De Bin, Sartori, and Severini (2015) and Pakel (2019), with the latter allowing for both time-series and cross-sectional dependence.

Our goal is to construct an IL that behaves like the target likelihood for the parameter of interest. The IL we construct to estimate $\theta$ is based on extending the approach in Severini (2007) to panel data models. Specifically, the maximum integrated likelihood estimator (MILE) that we propose is based on a certain data-dependent transformation of $\alpha_i$, called the "zero-score-expectation (ZSE)" transformation, which is used to construct an IL possessing desirable properties, irrespective of the weight-function used to integrate out the transformed nuisance parameter. The ZSE transformation ensures that, regardless of the weight-function, our IL is closer to the target likelihood, i.e., the Bartlett identities for it are closer to being satisfied, which has positive implications for estimation and inference.

The usual approach to determine whether a random function behaves like a genuine likelihood is to check if it satisfies the Bartlett identities, particularly score unbiasedness (the 1st Bartlett identity) and information unbiasedness (the 2nd Bartlett identity). In standard parametric panel data models, the profile likelihood satisfies both identities with error $O(1/T)$. The IL of AB and Lancaster satisfy the first identity with error $O(1/T^2)$, whereas the second holds with error $O(1/T)$, cf. Section 7.4. In contrast, our IL satisfies both identities with error $O(1/T^2)$, cf. Sections 7.3 and 7.4; i.e., unlike AB and Lancaster, our IL is simultaneously first-order score and first-order information unbiased. There is, therefore, an intuitive sense in which our IL improves upon those of AB and Lancaster. However, a rigorous proof that the distance (in some metric) between our IL and the target likelihood is smaller than the distance between the IL of AB or Lancaster and the target likelihood is beyond the scope of this paper. Note that Pace and Salvan (2006, Section 3.3) have shown that in cross-sectional models the Cox-Reid adjusted profile likelihood, which is not first-order information unbiased, and the modified profile likelihood of Barndorff-Nielsen, which is first-order information unbiased, both approximate the target likelihood to the same order. A Laplace approximation then suggests that the same is generally true for all related ILs as well. Consequently, it is the higher order terms that determine which IL better approximates the target likelihood. However, higher order IL calculations are beyond the scope of this paper.

Before we show how the ZSE transformed IL and the MILE are constructed, we briefly compare our paper with Severini (2007), AB, and Lancaster. There are no individual specific parameters in Severini (2007), who studies a pure cross-sectional model with a nuisance parameter whose dimension does not grow with $n$. In this context, Severini defines the concept of strong unrelatedness at the sample level and shows how to construct: (a) the ZSE transformation to create a nuisance parameter that is strongly unrelated to the parameter of interest, and, based on this transformation, (b) an integrated likelihood with desirable properties. Unlike us, Severini does not consider estimation or inference of the parameter of interest. Indeed,

none of the results obtained in our paper — including the definition of strong unrelatedness at the population level, the conditions under which the ZSE transformation and its inverse exist, rigorous derivations of the desirable properties of the ZSE based IL, the relationship between the ZSE transformation and the weight-functions of AB, the definition of the MILE and its application to various panel data models, the consistency and asymptotic normality of the MILE as $n, T \to \infty$, and its behavior in small samples — can be found in Severini (2007).

Compared to our approach, AB construct their IL differently. Namely, instead of transforming $\alpha_i$, they find data-dependent weight-functions to define their IL. The MILE has the same asymptotic distribution as $\hat{\theta}_{\text{AB}}$, even though $\hat{\theta}_{\text{AB}}$ is motivated and implemented differently. This is the sense in which our approach is complementary to that of AB (cf. Section 7.5 for more on this). On the other hand, we also extend the approach of AB as follows: (i) Our IL, unlike that of AB, is invariant to interest respecting transformations (Section 7.1). As a consequence, the MILE, unlike $\hat{\theta}_{\text{AB}}$, is also invariant to interest respecting transformations. (ii) Unlike the IL of AB, no special "weight-functions" are needed to construct our IL (Section 7.2). Indeed, a weight of unity, i.e., an "improper" weight-function, is sufficient for our IL to reduce both score and information bias. In contrast, the weight-functions proposed by AB are only guaranteed to reduce the score bias of their IL (Sections 7.3 and 7.4). Moreover, the results in Sections 7.3 and 7.4 can be used to characterize the conditions under which the AB weights can reduce both score and information bias (Section 7.5). (iii) We show the asymptotic normality of the MILE under a weaker rate condition than AB and Hahn and Kuersteiner (2011). Namely, they require $\lim_{n\to\infty} n/T_n \in (0, \infty)$, whereas we only require $\lim_{n\to\infty} n/T_n^3 = 0$ (Sections 8.2 and 9.2). Note that $\lim_{n\to\infty} n/T_n \in (0, \infty)$ implies that $n$ and $T_n$ grow at the same rate, whereas $\lim_{n\to\infty} n/T_n^3 = 0$, which implies that $T_n^3$ grows faster than $n$, allows for the possibility that $T_n$ itself can grow much slower than $n$. Hence, the smallness of $T$ relative to $n$ can be modeled more flexibly under the second condition. (iv) In finite samples, the MILE outperforms $\hat{\theta}_{\text{AB}}$, especially strikingly when $T$ is small (Section 10). Theoretical justification for this finding is given in Sections 8.3 and 10.

Our methodology also extends Lancaster's approach because the ZSE transformation can exist even when the IOT does not (Section 4). Hence, our approach is applicable to a wider class of models. Unlike Lancaster, our IL (and estimator) is invariant to interest respecting reparametrizations of the original likelihood and is simultaneously first-order score and information unbiased.

## 4. THE ZSE TRANSFORMATION

The fixed effects MLE of $\theta$ is inconsistent because estimation of $\alpha_i$ influences the estimation of $\theta$. We propose to fix this problem by transforming $\alpha_i$ — using the ZSE transformation of Severini (2007) defined subsequently — into a new "functional" nuisance parameter that

is "strongly unrelated" to $\theta$ (Definition 4.1).[1] This leads to a transformed likelihood, from which the new nuisance parameter is eliminated by integrating it out. This IL is then used to construct an estimator of $\theta$ having the desired properties. As mentioned earlier, unlike AB, the choice of weight-functions used to integrate out the nuisance parameter is not critical in our approach.

We begin by giving some intuition behind the ZSE transformation, which, loosely speaking, is a data-based bijective function $\alpha_i \mapsto g(\alpha_i)$ such that $g(\alpha^*_{iT}(\theta))$, the transformed population level MLE of $\alpha_i$, does not depend on $\theta$. How can such a mapping be constructed? Observe that the first-order condition (FOC) for $\alpha^*_{iT}(\theta)$ is $\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha^*_{iT}(\theta)); \theta_0, \alpha_{i0}] = 0$. Hence, if $g$ is defined as solving $\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \theta_0, g(\alpha_i))] = 0$ for each $\alpha_i$, then $g(\alpha^*_{iT}(\theta)) = \alpha_{i0}$, i.e., $g(\alpha^*_{iT}(\theta))$ does not depend on $\theta$ as desired. Although $g$ constructed in this manner depends generally on $\theta_0$ (which is unknown), a feasible version of $g$ can be obtained by replacing $\theta_0$ by a preliminary estimator, e.g., the fixed effects MLE, which is consistent as $n, T \to \infty$. We now make these notions precise.

Let $g_{iT\theta_0\theta} : (\mathfrak{a}, \mathfrak{b}) \to (\mathfrak{a}, \mathfrak{b})$ denote a function, which depends on $i, T, \theta_0, \theta$, such that for all $\alpha_i \in (\mathfrak{a}, \mathfrak{b})$, $g_{iT\theta_0\theta}(\alpha_i)$ satisfies the moment condition $\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \theta_0, g_{iT\theta_0\theta}(\alpha_i)] = 0$. Following Severini (2007, p. 532), we will refer to $\alpha \mapsto g_{iT\theta_0\theta}(\alpha)$ as the ZSE transformation and $\phi := g_{iT\theta_0\theta}(\alpha)$ as the ZSE (nuisance) parameter corresponding to $\alpha$ for a given $\theta$. The next result shows that the existence and uniqueness of $\alpha \mapsto g_{iT\theta_0\theta}(\alpha)$ and its inverse $\phi \mapsto h_{iT\theta_0\theta}(\phi)$, which is required to define the transformed likelihood in terms of the ZSE parameter, follows from the implicit function theorem.

**Lemma 4.1.** *Let $\mathbb{E}_0\ell^2_{iT\alpha}(\theta_0, \alpha_{i0}) > 0$. Then*: (i) *There exist open sets $\mathcal{B}_1 \ni \theta_0$, $\mathcal{D}_{1i} \ni \alpha_{i0}$, and $\mathcal{C}_{1i} \ni \alpha_{i0}$, such that, for each $(\theta, \alpha) \in \mathcal{B}_1 \times \mathcal{D}_{1i}$, there is a unique number $g_{iT\theta_0\theta}(\alpha) \in \mathcal{C}_{1i}$ satisfying $\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha); \theta_0, g_{iT\theta_0\theta}(\alpha)] = 0$.* (ii) *There exist open sets $\mathcal{B}_2 \ni \theta_0$, $\mathcal{D}_{2i} \ni \alpha_{i0}$, and $\mathcal{C}_{2i} \ni \alpha_{i0}$, such that, for each $(\theta, \phi) \in \mathcal{B}_2 \times \mathcal{D}_{2i}$, there is a unique number $h_{iT\theta_0\theta}(\phi) \in \mathcal{C}_{2i}$ satisfying $\mathbb{E}[\ell_{iT\alpha}(\theta, h_{iT\theta_0\theta}(\phi)); \theta_0, \phi] = 0$.* (iii) *There exists an open set $\mathcal{B} \subset \mathcal{B}_1 \cap \mathcal{B}_2$ with $\theta_0 \in \mathcal{B}$, and an open set $\mathcal{D}_i \subset \mathcal{D}_{1i} \cap \mathcal{D}_{2i}$ with $\alpha_{i0} \in \mathcal{D}_i$, such that, for each $\theta \in \mathcal{B}$, $h_{iT\theta_0\theta}$ is the inverse of $g_{iT\theta_0\theta}$ on $\mathcal{D}_i$, i.e., for each $\theta \in \mathcal{B}$, $h_{iT\theta_0\theta} \circ g_{iT\theta_0\theta}$ and $g_{iT\theta_0\theta} \circ h_{iT\theta_0\theta}$ are the identity map on $\mathcal{D}_i$.*

Lemma 4.1(ii) implies that the domain of the inverse ZSE transformation $h_{iT\theta_0\theta}$, denoted by $\mathcal{D}(h_{iT\theta_0\theta}) := \{\phi \in (\mathfrak{a}, \mathfrak{b}) :$ the equation $\mathbb{E}[\ell_{iT\alpha}(\theta, h); \theta_0, \phi] = 0$ can be solved for $h\}$, is a nonempty subset of $(\mathfrak{a}, \mathfrak{b})$ for each $\theta$ close enough to $\theta_0$. However, if $\ell_{iT}(\theta, \alpha_i)$ is sufficiently well behaved, then the argument in the next paragraph suggests that $\mathcal{D}(h_{iT\theta_0\theta}) = (\mathfrak{a}, \mathfrak{b})$ for

[1]Since $\alpha_i$ is individual specific, we are free to transform it provided the transformed value is also individual specific so that the interpretation of the model is not altered. Strong unrelatedness of the parameters has several important consequences. Cf. the discussion after Lemma 4.2, and Section 7, for more on this.

each $\theta_0, \theta \in \Theta$, i.e., the inverse ZSE transformation exists globally on $(\mathfrak{a}, \mathfrak{b})$. Indeed, in specific examples, such as those considered in Section 9, it is easy to see that, for each $\theta_0, \theta \in \Theta$, the inverse ZSE transformation $h_{iT\theta_0\theta}$ exists and $\mathcal{D}(h_{iT\theta_0\theta}) = (\mathfrak{a}, \mathfrak{b})$.

Let $(\theta, \phi) \in \mathcal{B}_2 \times \mathcal{D}_{2i}$. If the optimization problem $\max_{u \in \mathcal{D}_{2i}} \mathbb{E}[\ell_{iT}(\theta, u); \theta_0, \phi]$ has a unique solution then, by Lemma 4.1(ii), that solution must be $h_{iT\theta_0\theta}(\phi)$. In other words, $h_{iT\theta_0\theta}(\phi)$ is the population level MLE of $\alpha_i \in \mathcal{D}_{2i}$ when the true value of $(\theta, \alpha_i)$ is $(\theta_0, \phi) \in \mathcal{B}_2 \times \mathcal{D}_{2i}$. Extending this analogy, if the population level MLE of $\alpha_i \in (\mathfrak{a}, \mathfrak{b})$ exists for all true values of $(\theta, \alpha_i)$ in $\Theta \times (\mathfrak{a}, \mathfrak{b})$, i.e., if the optimization problem $\max_{u \in (\mathfrak{a}, \mathfrak{b})} \mathbb{E}[\ell_{iT}(\theta, u); \theta_0, \phi]$ has a unique solution for each $(\theta, \theta_0, \phi) \in \Theta \times \Theta \times (\mathfrak{a}, \mathfrak{b})$ — which, e.g., is the case if $u \mapsto \mathbb{E}[\ell_{iT}(\theta, u); \theta_0, \phi]$ is strictly concave on $(\mathfrak{a}, \mathfrak{b})$ for each $(\theta, \theta_0, \phi) \in \Theta \times \Theta \times (\mathfrak{a}, \mathfrak{b})$ — then, for each $\theta_0, \theta \in \Theta$, the inverse ZSE transformation $h_{iT\theta_0\theta}$ exists and $\mathcal{D}(h_{iT\theta_0\theta}) = (\mathfrak{a}, \mathfrak{b})$. In addition to providing a statistical interpretation to $h_{iT\theta_0\theta}$, this argument also helps explain why the inverse ZSE transformation can exist even when the IOT does not.[2]

The population level MLE interpretation of $h_{iT\theta_0\theta}$ justifies its global, i.e., for all $\theta_0, \theta \in \Theta$, existence as compared to Lemma 4.1, which only gives the conditions for its local, i.e., for $\theta$ close to $\theta_0$, existence. This is the motivation for Assumption C.1, which ensures that the ZSE transformation exists globally as a bijection from $(\mathfrak{a}, \mathfrak{b}) \to (\mathfrak{a}, \mathfrak{b})$. Under Assumption C.1, which strengthens Lemma 4.1, the inverse ZSE transformation $h_{iT\theta_0\theta}$ exists globally as a function from $(\mathfrak{a}, \mathfrak{b}) \to (\mathfrak{a}, \mathfrak{b})$, i.e., for each $(i, T, \theta_0, \theta) \in \mathbb{N} \times \mathbb{N} \times \Theta \times \Theta$, $\mathbb{E}[\ell_{iT\alpha}(\theta, h_{iT\theta_0\theta}(\phi)); \theta_0, \phi] = 0$ for $\phi \in (\mathfrak{a}, \mathfrak{b})$. The inverse ZSE transformation is used to construct $\tilde{L}^0_{iT}(\theta, \phi) := L_{iT}(\theta, h_{iT\theta_0\theta}(\phi))$, the infeasible (it depends upon $\theta_0$) ZSE transformed likelihood of $(\theta, \phi)$ for the $i$th individual. The usefulness of the corresponding loglikelihood $\tilde{\ell}^0_{iT}(\theta, \phi) := T^{-1} \log \tilde{L}^0_{iT}(\theta, \phi) = \ell_{iT}(\theta, h_{iT\theta_0\theta}(\phi))$ stems from a remarkable property of

$$\phi^*_{iT}(\theta) := \operatorname*{argmax}_{\phi \in (\mathfrak{a}, \mathfrak{b})} \mathbb{E}_0 \tilde{\ell}^0_{iT}(\theta, \phi), \qquad \theta \in \Theta,$$

the population level MLE of the ZSE parameter. This property, described in Lemma 4.2, is the reason why the ZSE transformation is useful. But we first need the following definition.

**Definition 4.1** (Strong unrelatedness)**.** *In the loglikelihood $\ell_{iT}(\theta, \alpha_i)$, the individual specific nuisance parameter $\alpha_i$ is said to be strongly unrelated at the population level to the parameter of interest $\theta$ if $\alpha^*_{iT}(\theta) = \alpha_{i0}$ for each $\theta \in \Theta$.*

In other words, $\alpha_i$ is strongly unrelated to $\theta$ at the population level if its population level MLE (for fixed $\theta$) does not depend on $\theta$. Severini (2007, p. 530) defines the strong unrelatedness

[2]The differential equations that define the IOT may not be solvable if $\dim(\theta) > 1$ (Cox and Reid, Section 2.3). Hence, the IOT is not guaranteed to exist if $\theta$ is a vector. However, as the inverse ZSE transformation is characterized differently, it can exist even when the IOT does not. For instance, although the IOT does not exist for an AR(1) model with covariates (Lancaster, 2002, Section 3.2), the calculations in Example 9.5 can be straightforwardly extended to show that $h_{iT\theta_0\theta}$ does exist in this model.

property in terms of estimators. Definition 4.1 is the analogous version in terms of population level parameters.

**Lemma 4.2.** *Under Assumption* C.1, $\theta \mapsto \phi^*_{iT}(\theta)$ *is constant on* $\Theta$. *In particular,* $\phi^*_{iT}(\theta) = \alpha_{i0}$ *for each* $\theta \in \Theta$.

Lemma 4.2 reveals that, in the infeasible ZSE transformed loglikelihood $\tilde{\ell}^0_{iT}(\theta, \phi)$, the ZSE parameter $\phi$ is strongly unrelated to the parameter of interest $\theta$ at the population level. In asymptotic expansions, strong unrelatedness of $\phi$ and $\theta$ allows $\phi^*_{iT}(\theta)$ to be replaced by $\alpha_{i0}$ without creating bias. For instance, this is used to justify Assumptions C.3(vi, vii), which are required to simplify the form of the individual log-IL (cf. (F.1)), and to show that it is insensitive to the choice of the weight-function (Section 7.2). The strong unrelatedness property also suggests that eliminating $\phi$ from $\tilde{\ell}^0_{iT}(\theta, \phi)$ will not affect the estimation of $\theta$. Indeed, as demonstrated subsequently, it is the strong unrelatedness of $\phi$ and $\theta$ that reduces both score and information bias for the ZSE transformed IL.

A consequence of $\phi$ and $\theta$ being strongly unrelated is that they are information orthogonal,[3] i.e., as shown in Appendix D,

$$\mathbb{E}_0 \nabla^2_{\theta\phi} \tilde{\ell}^0_{iT}(\theta_0, \alpha_{i0}) = 0, \tag{4.1}$$

where $\nabla^2_{ab} := \partial_b \circ \nabla_a$ and $\nabla^2_{\theta\phi} \tilde{\ell}^0_{iT}(\theta_0, \alpha_{i0}) := \nabla^2_{\theta\phi} \tilde{\ell}^0_{iT}(\theta, \phi)\big|_{\theta=\theta_0, \phi=\alpha_{i0}}$. This shows that the ZSE transformation makes $\theta$ and $\phi$ information orthogonal, even though it is not the IOT because it is characterized differently. Indeed, since a nuisance parameter can be information orthogonal to the parameter of interest without being strongly unrelated to it (cf. Example D.1), it follows that the ZSE transformation and the IOT are different objects.

## 5. THE MILE

We are now ready to define our estimator. Although the inverse ZSE transformation can be determined analytically in certain cases, cf. Examples 9.1, 9.2, and 9.5, it is typically obtained numerically as in Examples 9.3 and 9.4 (cf. Appendix A.1 for details). In principle, this is straightforward to do by fixing $\theta$ and then, for each given $\phi$, finding a number $h$ that numerically solves the equation $\mathbb{E}[\ell_{iT\alpha}(\theta, h); \theta_0, \phi] = 0$. In practice, however, $\theta_0$ is first replaced by a preliminary estimator, e.g., the fixed effects MLE $\tilde{\theta}$, which is consistent as $n, T \to \infty$. Then, given $\phi$, the equation $\mathbb{E}[\ell_{iT\alpha}(\theta, h); \tilde{\theta}, \phi] = 0$ is solved numerically for $h$. The solution is the function $\phi \mapsto h_{iT\tilde{\theta}\theta}(\phi)$, the estimator of the inverse ZSE transformation $\phi \mapsto h_{iT\theta_0\theta}(\phi)$.

[3]The ZSE transformation is not just another device to orthogonalize the parameters. Indeed, it achieves much more than simply orthogonalizing the parameters of the transformed likelihood, namely, it makes them strongly unrelated. It is the strong unrelatedness of the parameters (and not the fact that they are information orthogonal) that reduces both score and information bias for the ZSE transformed IL.

Let $\tilde{L}_{iT}(\theta,\phi) := L_{iT}(\theta, h_{iT\tilde{\theta}\theta}(\phi))$ denote the feasible version of $\tilde{L}^0_{iT}(\theta,\phi)$. Similarly, $\tilde{\ell}_{iT}(\theta,\phi) := T^{-1}\log \tilde{L}_{iT}(\theta,\phi) = \ell_{iT}(\theta, h_{iT\tilde{\theta}\theta}(\phi))$ is the feasible version of $\tilde{\ell}^0_{iT}(\theta,\phi)$. The feasible ZSE transformed IL for $\theta \in \Theta$ for the $i$th individual is defined to be

$$\bar{L}_{iT}(\theta) := \int_{(\mathfrak{a},\mathfrak{b})} \tilde{L}_{iT}(\theta,\phi)\pi_i(\phi)\,d\phi = \int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta, h_{iT\tilde{\theta}\theta}(\phi))\pi_i(\phi)\,d\phi, \tag{5.1}$$

where $\pi_i : (\mathfrak{a},\mathfrak{b}) \to (0,\infty)$ is a weight-function that does not depend on $\theta$, and it is assumed that the integral in (5.1) is finite for each $\theta \in \Theta$ (the necessary condition that $L_{iT}(\theta, h_{iT\tilde{\theta}\theta}(\cdot))$ exists for all $\theta \in \Theta$ follows from Assumption C.1). Unlike AB, the choice of $\pi_i$ here is not critical. Indeed, since $\nabla_\theta \bar{L}_{iT}(\theta)$ can be shown to be approximately independent of $\pi_i$ (Section 7.2), which is why we do not indicate the dependence of $\bar{L}_{iT}$ on $\pi_i$, it is perfectly acceptable to let $\pi_i := 1$, which is what we do in the examples (Section 9) and the simulations (Section 10).

Let $\bar{\ell}_{iT}(\theta) := T^{-1}\log \bar{L}_{iT}(\theta)$ denote the ZSE transformed log-IL for individual $i$. The MILE of $\theta$ is defined to be $\hat{\theta} := \operatorname{argmax}_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \bar{\ell}_{iT}(\theta)$, i.e.,[4]

$$\hat{\theta} := \operatorname*{argmax}_{\theta\in\Theta} n^{-1}\sum_{i=1}^n T^{-1}\log \int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta, h_{iT\tilde{\theta}\theta}(\phi))\pi_i(\phi)\,d\phi. \tag{5.2}$$

The definition of the MILE makes clear the difference between $\hat{\theta}$ and $\hat{\theta}_{\mathrm{AB}}$. Namely, we use the data-dependent ZSE transformation of the nuisance parameter to define our IL, whereas AB find a data-dependent weight-function for the nuisance parameter to define their IL.

The MILE can be iterated to remove its dependence on the preliminary estimator $\tilde{\theta}$, and perhaps even improve its finite sample properties: Once $\hat{\theta}$ becomes available, it is used to obtain $h_{iT\hat{\theta}\theta}$, which is then employed to recompute the MILE as defined in (5.2). This yields the single-iteration MILE $\hat{\theta}(1) := \operatorname{argmax}_{\theta\in\Theta} n^{-1}\sum_{i=1}^n T^{-1}\log\int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta, h_{iT\hat{\theta}\theta}(\phi))\pi_i(\phi)\,d\phi$. Repeating this process until convergence leads to the estimator denoted by $\hat{\theta}(\infty)$.

**Remark 5.1.** As with any interest respecting transformation (Section 7.1), the profile loglikelihood of the ZSE transformed loglikelihood $\tilde{\ell}_{iT}(\theta,\phi)$, given by $\tilde{\ell}_{iT}(\theta,\hat{\phi}_{iT\tilde{\theta}}(\theta))$, where

$$\hat{\phi}_{iT\tilde{\theta}}(\theta) := \operatorname*{argmax}_{\phi\in(\mathfrak{a},\mathfrak{b})} \ell_{iT}(\theta, h_{iT\tilde{\theta}\theta}(\phi)), \qquad \theta \in \Theta, \tag{5.3}$$

is identical to the profile loglikelihood in the original parameterization, i.e., cf. Appendix D,

$$\tilde{\ell}_{iT}(\theta,\hat{\phi}_{iT\tilde{\theta}}(\theta)) = \ell_{iT}(\theta,\hat{\alpha}_{iT}(\theta)), \qquad \theta \in \Theta. \tag{5.4}$$

[4]Although the ZSE transformed IL is constructed using the fixed effects MLE, whose asymptotic distribution is biased as $n,T \to \infty$, it is shown in Section 8.2 that the limiting distribution of the MILE is correctly centered as $n,T \to \infty$. Consequently, the MILE is not hindered by the fact that the preliminary estimator used to construct $h_{iT\tilde{\theta}\theta}$, namely, the fixed effects MLE, is asymptotically biased.

Consequently, in the definition of the MILE, integrating out the ZSE parameter, instead maximizing it out, is critical because the optimization problem $\operatorname{argmax}_{\theta\in\Theta}\sum_{i=1}^{n}\max_{\phi\in(\mathfrak{a},\mathfrak{b})}\tilde{\ell}_{iT}(\theta,\phi)$ simply yields the fixed effects MLE. □

## 6. APPROXIMATING THE ZSE TRANSFORMED IL

To show the properties of the ZSE transformed IL in Section 7, and to demonstrate the consistency and asymptotic normality of $\hat{\theta}$ in Section 8, we require the Laplace approximation of $\bar{\ell}_{iT}(\theta)$. Lemma 6.1, proved in Appendix E, provides a uniform (in $\theta$) Laplace approximation of $\bar{\ell}_{iT}(\theta)$. Henceforth, let $c_T := (2T)^{-1}\log(2\pi/T)$. Throughout this section, if not stated explicitly, limits are taken as $T\to\infty$.

**Lemma 6.1** (Laplace approximation)**.** *Let Assumptions* 2.1–C.2 *hold and* $\tilde{\theta}\in\mathcal{M}$*. Then,*

$$\bar{\ell}_{iT}(\theta) = c_T + \ell_{iT}(\theta,\hat{\alpha}_{iT}(\theta)) - \frac{1}{2T}\log(-\tilde{\ell}_{iT\phi\phi}(\theta,\hat{\phi}_{iT\tilde{\theta}}(\theta))) + \frac{1}{T}\log\pi_i(\hat{\phi}_{iT\tilde{\theta}}(\theta)) + R_{iT}(\theta) \quad (6.1)$$

*for* $\theta\in\Theta$*, where* $\tilde{\ell}_{iT\phi\phi}(\theta,\phi) := \partial_\phi^2\tilde{\ell}_{iT}(\theta,\phi)$ *and* $\sup_{\theta\in\Theta}|R_{iT}(\theta)| = O_{\mathrm{p}}(T^{-2})$ *as* $T\to\infty$*.*

Assumption C.2 strengthens the conditions in Kass, Tierney, and Kadane (1990) to ensure that $T^2R_{iT}(\theta)$ is, uniformly in $\theta$, bounded in probability. Uniformity in $\theta$ is needed to show the consistency of the MILE (Section 8.1). The event $\tilde{\theta}\in\mathcal{M}$ occurs w.p.a.1 as $n,T\to\infty$ because the fixed effects MLE is consistent as $n,T\to\infty$ (Assumption C.6).

It is clear from Lemma 6.1 that $\bar{\ell}_{iT}(\theta)$, modulo a constant and an $O_{\mathrm{p}}(T^{-2})$ remainder term, is the sum of the profile loglikelihood, a term $\log(-\tilde{\ell}_{iT\phi\phi}(\theta,\hat{\phi}_{iT\tilde{\theta}}(\theta)))$ that reflects how the ZSE transformation "additively corrects" the profile loglikelihood, and the weight-function $\log\pi_i(\hat{\phi}_{iT\tilde{\theta}}(\theta))$. It is the correction term $\log(-\tilde{\ell}_{iT\phi\phi}(\theta,\hat{\phi}_{iT\tilde{\theta}}(\theta)))$, which includes a contribution from the inverse ZSE transformation (cf. (E.29)), that causes the IL, hence, the MILE, to possess the desired properties. In contrast, the weight-function has no real effect on the MILE.

The correction term $\log(-\tilde{\ell}_{iT\phi\phi}(\theta,\hat{\phi}_{iT\tilde{\theta}}(\theta)))$ can be contrasted with the adjustment to the profile loglikelihood in Cox and Reid (1987, Equation 10). Whereas Cox and Reid use the IOT to transform the nuisance parameter, we use the ZSE transformation, which has the advantage, relative to the IOT, that it can exist even when the IOT does not, and that it is invariant to interest respecting reparametrizations (Section 7.1).

Lemma 6.1 has a useful corollary that can be used to gain intuition behind the form of the ZSE transformed IL. Cf. Example 9.3 for an illustration.

**Corollary 6.1.** *Under the assumptions of Lemma* 6.1*,*

$$\bar{L}_{iT}(\theta) = \sqrt{\frac{2\pi}{T}}\frac{L_{iT}(\theta,\hat{\alpha}_{iT}(\theta))}{\sqrt{-\ell_{iT\alpha\alpha}(\theta,\hat{\alpha}_{iT}(\theta))}}\frac{\pi_i(\hat{\phi}_{iT\tilde{\theta}}(\theta))}{|\partial_\phi h_{iT\tilde{\theta}\theta}(\hat{\phi}_{iT\tilde{\theta}}(\theta))|}(1+O_{\mathrm{p}}(T^{-1})), \qquad \theta\in\Theta,$$

*where the* $O_{\mathrm{p}}(T^{-1})$ *term holds uniformly in* $\theta\in\Theta$*.*

A closed-form expression for $\partial_\phi h_{iT\tilde{\theta}\theta}(\hat{\phi}_{iT\tilde{\theta}}(\theta))$ can be obtained from (I.4), although it is often easier to obtain $\partial_\phi h_{iT\tilde{\theta}\theta}(\hat{\phi}_{iT\tilde{\theta}}(\theta))$, and $\hat{\phi}_{iT\tilde{\theta}}(\theta)$, directly from the equation defining the inverse ZSE transformation; cf. Example 9.3 for details.

## 7. PROPERTIES OF THE ZSE TRANSFORMED IL

In this section, we show that the infeasible ZSE transformed IL, defined as $\bar{L}^0_{iT}(\theta) := \int_{(\mathfrak{a},\mathfrak{b})} \tilde{L}^0_{iT}(\theta,\phi)\pi_i(\phi)\,d\phi = \int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta, h_{iT\theta_0\theta}(\phi))\pi_i(\phi)\,d\phi$, and the corresponding log-infeasible-IL, given by $\bar{\ell}^0_{iT}(\theta) := T^{-1}\log \bar{L}^0_{iT}(\theta)$, possess some desirable properties, which provide the intuition behind why the MILE is robust to the choice of the weight-functions and behaves very well in finite samples. Specifically, it is shown that: (1) $\bar{L}^0_{iT}(\theta)$ is invariant to interest respecting reparametrizations of $L_{iT}(\theta,\alpha_i)$; (2) the weight-function $\pi_i$ is irrelevant in the sense that the mean and variance of the IL score are (approximately) independent of $\pi_i$; (3) the IL score is first-order unbiased (in some cases, e.g., the Neyman-Scott model in Example 9.1 and the linear AR(1) model in Example 9.5, the score of the IL defined with $\pi_i := 1$ can be exactly unbiased). (4) the information bias is also of the order $1/T^2$; and (5) there is a connection between the ZSE transformed IL and the IL of AB. This is used to characterize the conditions under which the AB weights can reduce both score and information bias. Throughout this section, limits are taken as $T \to \infty$.

7.1. **Invariance.** An interest respecting transformation, i.e., a map of the form $(\theta,\alpha_i) \mapsto (\theta, b(\alpha_i))$, where $b$ is a bijection from $(\mathfrak{a},\mathfrak{b}) \to (\mathfrak{a},\mathfrak{b})$, does not change the ZSE transformed likelihood. In other words, $L_{iT}(\theta,\cdot)$ and $L_{iT}(\theta, b^{-1}(\cdot))$ both yield the same $\tilde{L}^0_{iT}(\theta,\cdot)$. To see this, recall that if the true value of $(\theta,\alpha_i)$ is $(\theta_0,\phi)$, then the population level MLE of $\alpha_i$ using the likelihood $L_{iT}(\theta,\alpha_i)$ is $h_{iT\theta_0\theta}(\phi)$. Thus, by the equivariance of optimizers (Lemma D.1), the population level MLE of $\beta_i$ using the likelihood $L_{iT}(\theta, b^{-1}(\beta_i))$, when the true value of $(\theta,\beta_i)$ is $(\theta_0,\phi)$, is $b(h_{iT\theta_0\theta}(\phi))$. Hence, $\tilde{L}^0_{iT}(\theta,\phi)$ does not change, which implies that $\bar{L}^0_{iT}(\theta)$ remains invariant to interest respecting transformations. Since the same argument applies to the feasible integrated likelihood $\bar{L}_{iT}(\theta)$, it follows that the MILE remains invariant to interesting respecting transformations, a property not shared by the estimators of AB and Lancaster.

Henceforth, the second argument in functionals of $\ell_{iT}(\theta,\alpha_i)$ is omitted when $\alpha_i$ is evaluated at $\alpha^*_{iT}(\theta)$, the population level MLE of $\alpha_i$. E.g., we write the target loglikelihood as $\ell_{iT}(\theta) := \ell_{iT}(\theta,\alpha^*_{iT}(\theta))$, and $\ell_{iT\alpha}(\theta) := \ell_{iT\alpha}(\theta,\alpha^*_{iT}(\theta))$. Similarly, the second argument in functionals of $\tilde{\ell}^0_{iT}(\theta,\phi)$ is omitted when $\phi$ is evaluated at $\alpha_{i0}$, which, by Lemma 4.2, is the population level MLE of $\phi$ in the ZSE transformed likelihood. E.g., we write $\tilde{\ell}^0_{iT\phi}(\theta) := \tilde{\ell}^0_{iT\phi}(\theta,\phi)|_{\phi=\alpha_{i0}}$ and $\tilde{\ell}^0_{iT\phi\phi}(\theta) := \tilde{\ell}^0_{iT\phi\phi}(\theta,\phi)|_{\phi=\alpha_{i0}}$. Properties 2, 3, and 4, which involve taking expectations and derivatives of $\bar{\ell}^0_{iT}$, are shown under Assumption C.4. Such an assumption is also used in Hahn

and Newey (2004, p. 1303) and, implicitly, in AB. The derivatives of $\bar{\ell}^0_{iT}$, which are functions of the derivatives of $h_{iT\theta_0\theta}$, are well defined under Assumption C.13.

7.2. **Irrelevance of $\pi_i$.** Let $\hat{\phi}_{iT\theta_0}(\theta) = \operatorname{argmax}_{\phi\in(\mathfrak{a},\mathfrak{b})} \ell_{iT}(\theta, h_{iT\theta_0\theta}(\phi))$ denote the infeasible (it depends on $\theta_0$) sample analog of $\phi^*_{iT}(\theta)$. Since $\theta_0 \in \mathcal{M}$, by Lemma 6.1 we have

$$\bar{\ell}^0_{iT}(\theta) = c_T + \ell_{iT}(\theta, \hat{\alpha}_{iT}(\theta)) - \frac{1}{2T}\log(-\tilde{\ell}^0_{iT\phi\phi}(\theta, \hat{\phi}_{iT\theta_0}(\theta))) + \frac{1}{T}\log\pi_i(\hat{\phi}_{iT\theta_0}(\theta)) + R^0_{iT}(\theta), \quad (7.1)$$

where $\sup_{\theta\in\Theta}|R^0_{iT}(\theta)| = O_{\mathrm{p}}(T^{-2})$, and the remainder $R^0_{iT}$ is obtained from Lemma 6.1 upon replacing $\tilde{\theta}$ with $\theta_0$. Further expanding the terms $\log(-\tilde{\ell}^0_{iT\phi\phi}(\theta, \hat{\phi}_{iT\theta_0}(\theta)))$ and $\log\pi_i(\hat{\phi}_{iT\theta_0}(\theta))$, it is shown in Appendix F that[5]

$$\begin{aligned}\bar{\ell}^0_{iT}(\theta) = c_T + \ell_{iT}(\theta, \hat{\alpha}_{iT}(\theta)) - \frac{1}{2T}\log(-\mathbb{E}_0\tilde{\ell}^0_{iT\phi\phi}(\theta)) - \frac{1}{2T}C_{iT}(\theta) \\ + \frac{1}{T}\log\pi_i(\alpha_{i0}) - \dot{\pi}_i(\alpha_{i0})\frac{1}{T}P_{iT}(\theta) + O_{\mathrm{p}}(\frac{1}{T^2}), \quad \theta\in\Theta, \quad (7.2)\end{aligned}$$

where $C_{iT}(\theta)$ defined in (F.4) satisfies $\mathbb{E}_0 C_{iT}(\theta) = 0$ for each $\theta$, $P_{iT}(\theta)$ defined in (F.19) satisfies $\mathbb{E}_0 P_{iT}(\theta) = 0$ for each $\theta$ ($C_{iT}$ and $P_{iT}$ do not depend on $\pi_i$), and $\dot{\pi}_i(\phi) := \partial_\phi \log\pi_i(\phi)$.

Let $\bar{\ell}^0_{iT\theta}(\theta) := \nabla_\theta \bar{\ell}^0_{iT}(\theta)$ denote the score corresponding to $\bar{\ell}^0_{iT}(\theta)$. We demonstrate that $\mathbb{E}_0\bar{\ell}^0_{iT\theta}(\theta_0)$ and $\operatorname{var}_0\bar{\ell}^0_{iT\theta}(\theta_0)$ are approximately independent of $\pi_i$ in the sense that they do not depend on $\pi_i$, up to a term of order $1/T^2$. Indeed, (7.2) reveals that $\bar{\ell}^0_{iT\theta}(\theta)$ depends on $\pi_i$ through multiplication with $T^{-1}\nabla_\theta P_{iT}(\theta)$ and the derivative of remainder term. However, $\mathbb{E}_0\nabla_\theta P_{iT}(\theta_0) = 0$ by differentiating the identify $\mathbb{E}_0 P_{iT}(\theta) = 0$, cf. (F.20), and the expectation ($\mathbb{E}_0$) of the derivative of remainder term is $O_{\mathrm{p}}(T^{-2})$ by Assumption C.4. It follows that $\mathbb{E}_0\bar{\ell}^0_{iT\theta}(\theta_0)$ does not depend on $\pi_i$, up to a term of order $1/T^2$. Furthermore, it is shown in Appendix F (cf. p. 75 of the supplement) that elements of the matrix $\operatorname{var}_0\bar{\ell}^0_{iT\theta}(\theta_0)$ also do not depend on $\pi_i$, up to a term of order $1/T^2$.

**Remark 7.1** (Why $n/T^3 \to 0$?)**.** The fact that the mean and variance of the IL score are independent of the weight-function $\pi_i$ up to a term of order $1/T^2$, implies that $\pi_i$ will not affect the limiting distribution of the IL score, hence, that of the MILE, if $n/T^3 \to 0$. Indeed, any term on the right-hand side of (7.2) whose derivatives with respect to $\theta$ have means and variances (evaluated at $\theta_0$) of order $1/T^2$ or smaller, will not affect the limiting distribution of the MILE if $n/T^3 \to 0$. To see the intuition behind this claim, without being distracted by the presence of $\tilde{\theta}$ in (5.2), consider the infeasible MILE $\hat{\theta}^0 := \operatorname{argmax}_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \bar{\ell}^0_{iT}(\theta)$, which satisfies the FOC $\sum_{i=1}^n \bar{\ell}^0_{iT\theta}(\hat{\theta}^0) = 0$. Expanding each coordinate of $\bar{\ell}^0_{iT\theta}(\hat{\theta}^0)$ about $\theta_0$, we

[5]Equation (7.2) uses the fact that the ZSE parameter is strongly unrelated to $\theta$ at the population level. Indeed, without the strong unrelatedness property, the $O_{\mathrm{p}}(T^{-2})$ term in (7.2) would only be $O_{\mathrm{p}}(T^{-1})$; cf. the proof of (7.2) and Footnote 39 in the supplementary material.

have that $\sqrt{nT}(\hat{\theta}^0 - \theta_0) = (-\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\theta\theta}\bar{\ell}^0_{iT}(\bar{\theta}))^{-1}\sqrt{\frac{T}{n}}\sum_{i=1}^{n}\bar{\ell}^0_{iT\theta}(\theta_0)$, where $\bar{\theta}$ lies between $\hat{\theta}^0$ and $\theta_0$.[6] Hence, terms on the right-hand side of (7.2) whose derivatives have means and variances of order $1/T^2$ or smaller are asymptotically negligible and, thus, do not affect the limiting distribution of the MILE if $n/T^3 \to 0$. □

7.3. **First-order score unbiasedness.** It is shown in Lemma F.2 that the score of the infeasible ZSE transformed IL is first-order unbiased. Namely, for each $i$, the score of the infeasible ZSE transformed IL satisfies the first Bartlett identity up to a term of order $T^{-2}$, i.e.,

$$\mathbb{E}_0\bar{\ell}^0_{iT\theta}(\theta_0) = O_{\mathrm{p}}(T^{-2}). \tag{7.3}$$

To get some intuition behind this result, note that[7]

$$\mathbb{E}_0\bar{\ell}^0_{iT}(\theta) \overset{(7.2)}{=} c_T + \mathbb{E}_0\ell_{iT}(\theta, \hat{\alpha}_{iT}(\theta)) + B_{iT1}(\theta) + T^{-1}\log\pi_i(\alpha_{i0}) + O_{\mathrm{p}}(T^{-2}), \quad \theta \in \Theta,$$

where the $C_{iT}(\theta)$ and $P_{iT}(\theta)$ terms drop out because they have mean ($\mathbb{E}_0$) zero, $B_{iT1}(\theta)$ is defined in the proof of (7.3), and taking the expectation ($\mathbb{E}_0$) of the $O_{\mathrm{p}}(T^{-2})$ remainder does not change its rate (Assumption C.4). Differentiating with respect to $\theta$, and using that $\nabla_\theta\mathbb{E}_0\bar{\ell}^0_{iT}(\theta) = \mathbb{E}_0\bar{\ell}^0_{iT\theta}(\theta)$, it can be seen that the bias of the IL score has two sources: (a) the bias of the profile likelihood score, i.e., $\nabla_\theta\mathbb{E}_0\ell_{iT}(\theta, \hat{\alpha}_{iT}(\theta))$; and (b) the gradient of the correction term, i.e., $\nabla_\theta B_{iT1}(\theta)$. It is shown in the proof of (7.3) that, evaluated at $\theta_0$, (a) and (b) cancel each other out modulo a term of order $T^{-2}$, thereby leaving $\mathbb{E}_0\bar{\ell}^0_{iT\theta}(\theta_0)$ of order $T^{-2}$.

7.4. **First-order information unbiasedness.** It is shown in Lemma F.3 that, for each $i$, the information equality for the infeasible ZSE transformed IL holds up to an $O_{\mathrm{p}}(T^{-2})$ term, i.e.,[8]

$$T\mathbb{E}_0\nabla_\theta\bar{\ell}^0_{iT}(\theta_0)\partial_\theta\bar{\ell}^0_{iT}(\theta_0) + \mathbb{E}_0\nabla^2_{\theta\theta}\bar{\ell}^0_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-2}). \tag{7.4}$$

In other words, for each individual, the information of the infeasible ZSE transformed IL is first-order unbiased. By contrast, as demonstrated in Appendix F (cf. page 84 of the supplementary material), the weight-functions of AB do not, in general, reduce the information bias. The result that Lancaster's IL does not generally reduce the information bias follows from DiCiccio, Martin, Stern, and Young (1996, Section 3.1, p. 194).

Information unbiasedness, i.e., the 2nd Bartlett identity, is useful for at least two important reasons. First, information unbiasedness guarantees the asymptotic efficiency of the MLE

[6]Since the expansion is coordinatewise, the mean value $\bar{\theta}$ differs across the coordinates. However, in order to avoid overloading the notation, this dependence is suppressed throughout the paper.

[7]The expression for $\mathbb{E}_0\bar{\ell}^0_{iT}(\theta)$ uses the fact that the ZSE parameter and $\theta$ are strongly unrelated at the population level (because it follows from (F.1) and (F.3), and the latter are based on this property).

[8]The proof of (7.4) also uses the fact that the ZSE parameter and $\theta$ are strongly unrelated at the population level (because it also requires (F.1) and (F.3), cf. (F.50)).

(which is an $M$-estimator) by ensuring that the "sandwich form" of its asymptotic variance equals the inverse of the Fisher information. Since the infeasible ZSE transformed IL is always first-order information unbiased (read: close to being information unbiased), whereas the IL of AB or Lancaster is not always first-order information unbiased (read: not close to being information unbiased), this suggests that in finite samples the variance of the MILE may be smaller than the variance of $\hat{\theta}_{\text{AB}}$, at least when $T$ is small.[9] Second, information unbiasedness is useful because it is fundamental for testing model specification. In panel data models with fixed effects, it is not possible to directly test the specification of $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$ because $\alpha_{i0}$ is unobserved. Instead, once the individual specific nuisance parameters have been eliminated, an information matrix (IM) test can be applied to the ZSE transformed IL $n^{-1}\sum_{i=1}^{n}\bar{\ell}_{iT}(\theta)$. Equation 7.4 suggests that an IM test based on the ZSE transformed IL should have better size properties than an IM test based on the IL of AB or Lancaster, because the information equality is easier to reject using the IL of AB or Lancaster due to the fact that they have higher information bias, namely, of order $O_{\text{p}}(T^{-1})$. However, investigating misspecification issues is beyond the scope of our paper.

The fact that the score and information of the ZSE transformed IL are both first-order unbiased suggests (as mentioned earlier in Section 3) that our IL is, at least in an intuitive sense, closer to the target likelihood than the IL of AB or Lancaster.[10]

7.5. **Relationship between the ZSE transformation and AB weight-functions.** Unlike the weight-functions in equations (12) and (14) of AB that only reduce score bias, the derivative of the ZSE transformation can be characterized as a weight-function that reduces both score and information bias. To see this, observe that

$$\pi_i := 1 \implies \bar{L}^0_{iT}(\theta) = \int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta, h_{iT\theta_0\theta}(\phi))\, d\phi = \int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta,\alpha)|\partial_\alpha g_{iT\theta_0\theta}(\alpha)|\, d\alpha$$

by a change of variables. In other words, $\bar{L}^0_{iT}(\theta)$ (with $\pi_i := 1$) can be interpreted as an IL based on $L_{iT}(\theta,\cdot)$ using the weight-function $|\partial_\alpha g_{iT\theta_0\theta}(\cdot)|$. Since we have already shown that $\bar{L}^0_{iT}(\theta)$ is first-order score and information unbiased, it follows that $|\partial_\alpha g_{iT\theta_0\theta}(\cdot)|$ is a weight-function for $L_{iT}(\theta,\cdot)$ that reduces both score and information bias.

[9]This is best seen from Table 2, which reveals that for the probit model — where the infeasible ZSE transformed IL is first-order information unbiased, but the IL of AB and Lancaster are not — the variance of the MILE is smaller than the variance of $\hat{\theta}_{\text{AB}}$ and Lancaster's estimator for $T = 5$. Theoretical justification for this is given in Section 10.2.

[10]The profile likelihood is, in general, also not first-order information unbiased (DiCiccio et al., p. 190). For instance, it can be shown that in logit and probit models the information bias of the profile likelihood is of the order $T^{-1}$, whereas in the AR(1) model of Example 9.5 it is of order $T^{-2}$ (Appendix H, p. 123). This suggests that our IL is closer to the target likelihood than the profile likelihood as well.

Since the weight-function $|\partial_\alpha g_{iT\theta_0\theta}(\cdot)|$ reduces both score and information bias, we could have used it to define the MILE as $\operatorname{argmax}_\theta n^{-1}\sum_{i=1}^n T^{-1}\log\int_{(\mathfrak{a},\mathfrak{b})} L_{iT}(\theta,\alpha)|\partial_\alpha g_{iT\tilde{\theta}\theta}(\alpha)|\,d\alpha$. However, given that $g_{iT\tilde{\theta}\theta}$ is typically determined numerically, its derivative $\partial_\alpha g_{iT\tilde{\theta}\theta}$ will have to be obtained numerically as well and this will increase the computational complexity of this estimator. In contrast, $\hat{\theta}$ (with $\pi_i := 1$) does not require the computation of derivatives of $h_{iT\theta_0\theta}$ and, hence, it is more stable to implement than the former. Of course, if $g_{iT\theta_0\theta}$ and $h_{iT\theta_0\theta}$ are known analytically, as in Example 9.5, then the two estimators will coincide.

Note that $|\partial_\alpha g_{iT\theta_0\theta}(\cdot)|$ does not coincide with any of the specific members of the "robust" class of weight-functions emphasized in AB and used in their simulations. Indeed, using (I.4), it is straightforward to show that

$$\partial_\alpha g_{iT\theta_0\theta}(\alpha) = -\frac{1}{T}\frac{\mathbb{E}[\ell_{iT\alpha\alpha}(\theta,\alpha);\theta_0,g_{iT\theta_0\theta}(\alpha)]}{\mathbb{E}[\ell_{iT\alpha}(\theta,\alpha)\ell_{iT\alpha}(\theta_0,g_{iT\theta_0\theta}(\alpha));\theta_0,g_{iT\theta_0\theta}(\alpha)]}. \tag{7.5}$$

Comparing this expression with the weight-functions given in equations (12) and (14) of AB, it is clear that $|\partial_\alpha g_{iT\theta_0\theta}(\cdot)|$ is different from both of them.

A useful consequence of our results on information bias reduction is that we can provide necessary and sufficient conditions for the weight-functions of AB to be both first-order score and information bias reducing. Let $w_i(\theta,\cdot)$ denote the infeasible weight-function in the IL of AB. Then, as shown in Appendix F:

**Proposition 7.1.** *$w_i(\theta,\cdot)$ is first-order score and information bias reducing if and only if*

(i) $\nabla_\theta[\log w_i(\theta,\alpha^*_{iT}(\theta)) - \log(\frac{-\mathbb{E}_0\ell_{iT\alpha\alpha}(\theta)}{T\mathbb{E}_0\ell_{iT\alpha}(\theta)\ell_{iT\alpha}})]\big|_{\theta=\theta_0} = O_{\mathrm{p}}(T^{-1})$ *and*

(ii) $\nabla^2_{\theta\theta}[\log w_i(\theta,\alpha^*_{iT}(\theta)) - \log(\frac{-\mathbb{E}_0\ell_{iT\alpha\alpha}(\theta)}{T\mathbb{E}_0\ell_{iT\alpha}(\theta)\ell_{iT\alpha}})]\big|_{\theta=\theta_0} = O_{\mathrm{p}}(T^{-1})$.

(i) ensures that $w_i$ eliminates the first-order score bias, and (ii) ensures that $w_i$ eliminates the first-order information bias. (i) is similar, but not the same, to the condition in AB (Theorem 2). The class of weight-functions satisfying (i) and (ii) is not empty. Indeed, using the fact that $g_{iT\theta_0\theta}(\alpha^*_{iT}(\theta)) = \alpha_{i0}$, which follows from (D.9) and the strong unrelatedness of the ZSE parameter and $\theta$ at the population level (Lemma 4.2), it is straightforward to verify that $\partial_\alpha g_{iT\theta_0\theta}$ in (7.5) satisfies both (i) and (ii) exactly, i.e., with their $O_{\mathrm{p}}(T^{-1})$ terms replaced by zero. In fact, any weight-function of the form $m_{iT}(\theta,\cdot)\partial_\alpha g_{iT\theta_0\theta}(\cdot)$, where $m_{iT}$ is a positive data based function satisfying $\mathbb{E}_0\nabla_\theta \log m_{iT}(\theta,\alpha^*_{iT}(\theta)) = O_{\mathrm{p}}(T^{-1})$ and $\mathbb{E}_0\nabla^2_{\theta\theta}\log m_{iT}(\theta,\alpha^*_{iT}(\theta)) = O_{\mathrm{p}}(T^{-1})$, will satisfy (i) and (ii).

## 8. ASYMPTOTIC PROPERTIES

In this section, we show that the MILE is consistent (Section 8.1), and asymptotically normal with correct centering (Section 8.2). These results are used to compare the asymptotic

behavior of the likelihood ratio statistics constructed using the ZSE transformed IL and the IL of AB (Section 8.3). Let $\bar{\ell}_{\cdot T}(\theta) := n^{-1}\sum_{i=1}^{n}\bar{\ell}_{iT}(\theta)$ denote the average integrated loglikelihood, and $Q^*_{\cdot T}(\theta) := n^{-1}\sum_{i=1}^{n}\ell_{iT}(\theta,\alpha^*_{iT}(\theta))$ the average target loglikelihood, for the entire sample. Throughout this section, even when not mentioned explicitly, limits are taken as $n, T_n \to \infty$.

8.1. **Consistency.** It is shown in Appendix G that $\hat{\theta}$ is consistent for $\theta_0$ without imposing a rate on $T_n$, i.e., $T_n$ can grow faster or slower than $n$.

**Theorem 8.1.** *Let Assumptions* 2.1–C.2 *and* C.5 *hold. Then,* $\hat{\theta} \xrightarrow{\mathrm{p}} \theta_0$ *as* $n, T_n \to \infty$.

Assumption C.5 is standard in the literature in consistency arguments. The first few conditions in Assumption C.5, e.g., compactness of $\Theta$, boundedness of $(\mathfrak{a}, \mathfrak{b})$, well behavior of the remainder term in Lemma 6.1, ensure that $\sup_{\theta\in\Theta}|\bar{\ell}_{\cdot T_n}(\theta) - \mathbb{E}Q^*_{\cdot T_n}(\theta)| = o_{\mathrm{p}}(1)$. The remaining conditions in Assumption C.5, namely, the uniform convergence of $Q^*_{\cdot T_n}(\cdot) - \mathbb{E}Q^*_{\cdot T_n}(\cdot)$ on $\Theta$ and the identification of $\theta_0$ as the well-separated maximum of $\mathbb{E}Q^*_{\cdot T_n}(\cdot)$ for all $n, T_n$ sufficiently large, then lead to a straightforward proof of the consistency of $\hat{\theta}$.

8.2. **Normality.** It is shown in Appendix G that, for each $i$, the information equality holds for the target likelihood, and that $F_{iT} := T\mathbb{E}_0[\nabla_\theta \ell_{iT}(\theta_0)\partial_\theta \ell_{iT}(\theta_0)]$ is the partial information for $\theta$ in the presence of $\alpha_i$, i.e.,

$$\begin{gathered} T\mathbb{E}_0[\nabla_\theta \ell_{iT}(\theta_0)\partial_\theta \ell_{iT}(\theta_0)] + \mathbb{E}_0\nabla^2_{\theta\theta}\ell_{iT}(\theta_0) = 0 \\ F_{iT} = F_{iT\theta\theta} - \frac{F'_{iT\alpha\theta}F_{iT\alpha\theta}}{F_{iT\alpha\alpha}}, \end{gathered} \tag{8.1}$$

where $F_{iT\theta\theta} := -\mathbb{E}_0\ell_{iT\theta\theta}$, $F_{iT\alpha\theta} := -\mathbb{E}_0\ell_{iT\alpha\theta}$, and $F_{iT\alpha\alpha} := -\mathbb{E}_0\ell_{iT\alpha\alpha}$. The following result, where $F := \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbb{E}F_{iT_n} \overset{\text{Ass. 2.3(i)}}{=} \lim_{n\to\infty}\mathbb{E}F_{1T_n} = \lim_{T\to\infty}\mathbb{E}F_{1T}$ denotes the limit of the average expected partial information for $\theta$, is proved in Appendix G.

**Theorem 8.2.** *Let Assumptions* 2.1–C.3 *and* C.5–C.12 *hold. Then,* $\sqrt{nT_n}(\hat{\theta}-\theta_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, F^{-1})$ *as* $n, T_n \to \infty$ *and* $n/T_n^3 \to 0$.

Assumptions C.3 and C.5–C.12 collectively ensure that, in a series of approximations, replacing $\nabla_\theta\bar{\ell}_{iT_n}, \nabla^2_{\theta\theta}\bar{\ell}_{iT_n}$ in the FOC of the MILE with $\nabla_\theta\bar{\ell}^0_{iT_n}, \nabla^2_{\theta\theta}\bar{\ell}^0_{iT_n}$, and the latter with their target loglikelihood versions $\nabla_\theta\ell_{iT_n}(\theta_0), \nabla^2_{\theta\theta}\ell_{iT_n}(\theta_0)$, does not affect the limiting distribution of $\sqrt{nT_n}(\hat{\theta}-\theta_0)$, i.e., the MILE has the same asymptotic distribution as the maximizer of the target likelihood for the entire sample. Assumption C.3 is used to simplify the right-hand side of (7.1), and Assumption C.10 ensures that the infeasible IL influence function $\sqrt{T_n/n}\sum_{i=1}^{n}\bar{\ell}^0_{iT_n\theta}(\theta_0)$ can be replaced by the target likelihood influence function $\sqrt{T_n/n}\sum_{i=1}^{n}\nabla_\theta\ell_{iT_n}(\theta_0)$ by incurring an $O_{\mathrm{p}}(T_n^{-1/2})$ term. Assumption C.6 is maintained to ensure that the fixed effects MLE is consistent. Assumption C.7 ensures that the inverse ZSE transformation is well behaved so that the ZSE transformed IL is well approximated by the target likelihood. Assumption C.8

ensures that twice differentiating the remainder in the Laplace approximation of the ZSE transformed IL does not alter its rate of convergence, and Assumption C.9 guarantees the same when the derivatives are averaged across the individuals. Assumption C.11 enables the application of a central limit theorem (CLT) for triangular arrays, and Assumption C.12 allows us to handle the hessian term in the FOC of the MILE. In Appendix C.6, we demonstrate how these assumptions can be verified for the panel logit model in Example 9.3.

The standard error of $\hat{\theta}$ is easily obtained from the hessian of the integrated loglikelihood. Indeed, the proof of (G.8) reveals that $-n^{-1}\sum_{i=1}^{n}\nabla^2_{\theta\theta}\bar{\ell}_{iT_n}(\hat{\theta}) \xrightarrow{\text{p}} F$ as $n, T_n \to \infty$. Hence, the variance of $\hat{\theta}$ is consistently estimated by $\hat{F}^{-1}$ with $\hat{F} := -n^{-1}\sum_{i=1}^{n}\nabla^2_{\theta\theta}\bar{\ell}_{iT_n}(\hat{\theta})$.

8.3. **Inference.** Since the ZSE transformed IL behaves like the target likelihood, inference for $\theta_0$ can be based on the likelihood ratio (LR) statistic $\text{LR}_{nT}(\theta) := 2nT[\bar{\ell}_{\cdot T}(\hat{\theta}) - \bar{\ell}_{\cdot T}(\theta)]$, $\theta \in \Theta$, constructed using the MILE. It is shown in Appendix G that, under the conditions maintained in Theorem 8.2,

$$\text{LR}_{nT_n}(\theta_0) \xrightarrow{\text{d}} \chi^2_{\dim(\theta_0)} \qquad \text{as } n, T_n \to \infty. \tag{8.2}$$

This result can be used to test hypotheses and construct confidence regions for $\theta_0$. E.g., (8.2) can be used to show that the lower-level random set $\{\theta \in \Theta : \text{LR}_{nT_n}(\theta) \leq k_\tau\}$, where $\tau \in (0,1)$ and $k_\tau$ denotes the $1-\tau$ quantile of a $\chi^2_{\dim(\theta_0)}$ random variable, is a confidence region for $\theta_0$ whose coverage probability approaches $1-\tau$ as $n, T_n \to \infty$.

Schumann, Severini, and Tripathi (2020), henceforth SST, show that the LR statistic based on the ZSE transformed IL, which is first-order score and information unbiased, satisfies

$$\mathbb{E}_0\text{LR}_{nT}(\theta_0) = \dim(\theta_0) + O_{\text{p}}(\frac{n}{T^3}) + O_{\text{p}}(\frac{1}{T^2}) \qquad \text{as } n, T \to \infty, \tag{8.3}$$

where, abusing notation, $\mathbb{E}_0$ now denotes expectation with respect to $\prod_{i=1}^{n} f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$, the joint density of outcomes, given the explanatory variables, for the entire sample. In contrast, $\text{LR}^{\text{AB}}_{nT}(\cdot)$, the LR statistic based on $\hat{\theta}_{\text{AB}}$ and the IL of AB, which is only first-order score (but not information) unbiased, satisfies

$$\mathbb{E}_0\text{LR}^{\text{AB}}_{nT}(\theta_0) = \dim(\theta_0) + O_{\text{p}}(\frac{n}{T^3}) + O_{\text{p}}(\frac{1}{T}) \qquad \text{as } n, T \to \infty. \tag{8.4}$$

The right-hand side of (8.3) contains the remainder $O_{\text{p}}(T^{-2})$, whereas the right-hand side of (8.4) contains the remainder $O_{\text{p}}(T^{-1})$. As explained in SST, this is because the ZSE transformed IL is first-order information unbiased, whereas the IL of AB is not. It follows that $\mathbb{E}_0\text{LR}_{nT}(\theta_0)$ is closer to its limiting value than $\mathbb{E}_0\text{LR}^{\text{AB}}_{nT}(\theta_0)$ due to the first-order information unbiasedness of the ZSE transformed IL. The fact that the ZSE transformed IL is first-order information unbiased thus helps explain why estimators that may have a similar mean squared error (MSE) as the MILE can do worse in terms of the empirical coverage of the corresponding LR statistics when $T$ is small; cf. Section 10 for more on this.

The performance of the LR statistic worsens for likelihoods that are neither first-order score, nor first-order information, unbiased. For instance, SST show that $\mathrm{LR}^{\mathrm{p}}_{nT}(\cdot)$, the LR statistic based on the fixed effects MLE and the profile likelihood, which is neither first-order score nor first-order information unbiased, satisfies

$$\mathbb{E}_0\mathrm{LR}^{\mathrm{p}}_{nT}(\theta_0) = \dim(\theta_0) + O_{\mathrm{p}}(\frac{n}{T}) + O_{\mathrm{p}}(\frac{1}{T}) \qquad \text{as } n, T \to \infty. \tag{8.5}$$

Asymptotic unbiasedness of $\mathrm{LR}^{\mathrm{p}}_{nT}(\theta_0)$ requires $T$ to grow faster than $n$, which explains why the empirical size of the LR statistic based on the profile likelihood can be far from its nominal level in short panels, as is evident from the simulation results in Section 10.

## 9. EXAMPLES

We now demonstrate how the ZSE approach works in some familiar settings. Henceforth, $\Lambda(u) := e^u/(1+e^u)$, $u \in \mathbb{R}$, is the logistic cdf, $\lambda(u) := d\Lambda(u)/du$ the logistic density, $\mathfrak{N}$ denotes the standard normal cdf, and $\mathfrak{n}(u) := d\mathfrak{N}(u)/du$ the standard normal density. Throughout this section, the ZSE transformed IL is constructed using $\pi_i := 1$ and $(\mathfrak{a}, \mathfrak{b}) = \mathbb{R}$.

### 9.1. **Static models.**

The static models we consider are the Neyman-Scott model, panel poisson, panel logit, and panel probit.

**Example 9.1** (Example B.1 contd.)**.** In the Neyman-Scott model, the parameter of interest is $\theta := \sigma^2$ and the likelihood for the $i$th individual is $L_{iT}(\theta, \alpha_i) = (2\pi)^{-T/2}\theta^{-T/2}e^{-\sum_{t=1}^T (Y_{it}-\alpha_i)^2/2\theta}$. Hence, $\ell_{iT\alpha}(\theta, \alpha_i) = \sum_{t=1}^T (Y_{it} - \alpha_i)/T\theta$ and

$$\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \check{\theta}, \check{\alpha}] = \frac{1}{T\theta}\sum_{t=1}^T (\mathbb{E}[Y_{it}; \check{\theta}, \check{\alpha}] - \alpha_i) = \frac{\check{\alpha} - \alpha_i}{\theta},$$

which implies that the ZSE transformation and its inverse are the identity mapping on $\mathbb{R}$, i.e., $g_{iT\theta_0\theta}(\alpha_i) = \alpha_i$ and $h_{iT\theta_0\theta}(\phi) = \phi$. Thus, $\tilde{L}_{iT}(\theta, \phi) = L_{iT}(\theta, \phi)$ because $h_{iT\theta_0\theta}$ does not depend upon $\theta_0$. Hence, letting $\bar{Y}_{i\cdot} := \sum_{t=1}^T Y_{it}/T$, the ZSE transformed IL for the $i$th individual is given by (modulo a factor that does not depend on $\theta$)

$$\begin{aligned}\bar{L}_{iT}(\theta) = \int_{\mathbb{R}} \tilde{L}_{iT}(\theta, \phi)\, d\phi &= \int_{\mathbb{R}} L_{iT}(\theta, \phi)\, d\phi \\ &= (2\pi)^{-(T-1)/2}T^{-1/2}\theta^{-(T-1)/2}e^{-\sum_{t=1}^T (Y_{it}-\bar{Y}_{i\cdot})^2/2\theta} \\ &\propto \frac{\mathrm{pdf}_{\mathcal{Y}_{iT}|\alpha_i;\theta}(\mathcal{Y}_{iT})}{\mathrm{pdf}_{\sum_{t=1}^T Y_{it}|\alpha_i;\theta}(\sum_{t=1}^T Y_{it})}.\end{aligned}$$

Thus, $\bar{L}_{iT}(\theta)$ is equal to the conditional likelihood, which does not depend upon $\alpha_i$ because $\sum_{t=1}^T Y_{it}$ is sufficient for $\alpha_i$. Therefore, the MILE coincides with the conditional maximum

likelihood estimator (CMLE), i.e.,

$$\hat{\sigma}^2 = \frac{1}{n(T-1)} \sum_{i=1}^{n} \sum_{t=1}^{T} (Y_{it} - \bar{Y}_{i\cdot})^2.$$

As shown in Appendix H, as $n \to \infty$ but $T$ is held fixed,

$$\begin{aligned} \hat{\sigma}^2 - \sigma_0^2 &= O_{\mathrm{p}}(\sqrt{\frac{1}{nT}}) + O_{\mathrm{p}}(\sqrt{\frac{1}{nT^2}}) \\ \sqrt{n(T-1)}(\hat{\sigma}^2 - \sigma_0^2) &\xrightarrow{\mathrm{d}} \mathrm{N}(0, 2\sigma_0^4). \end{aligned} \tag{9.1}$$

Hence, unlike the fixed effects MLE, the MILE is consistent as $n \to \infty$, $T$-fixed. The asymptotic normality of the MILE, as $n \to \infty$ and $T$ is held fixed, is also a much more robust result than the one demonstrated for the fixed effects MLE in (B.2). Letting $\bar{\ell}_{iT\theta}(\theta) := \nabla_\theta \bar{\ell}_{iT}(\theta)$, note that

$$\bar{\ell}_{iT\theta}(\theta) = -\frac{(T-1)}{T} \frac{1}{2\theta^2} [\theta - \frac{1}{T-1} \sum_{t=1}^{T} (Y_{it} - \bar{Y}_{i\cdot})^2] \implies \mathbb{E}_0 \bar{\ell}_{iT\theta}(\theta_0) = 0.$$

Therefore, the individual integrated loglikelihood scores are exactly unbiased for zero. Hence, the MILE is just the generalized method of moments (GMM) estimator of $\sigma^2$ based on the moment condition $\mathbb{E}_0[\theta_0 - \frac{1}{T-1} \sum_{t=1}^{T} (Y_{it} - \bar{Y}_{i\cdot})^2] = 0$, and the limiting distribution in (9.1) can also be obtained by applying GMM theory to this moment condition. □

**Example 9.2** (Panel poisson)**.** Let $Y_{it} \big| \mathcal{X}_{iT}, \alpha_{i0}; \theta_0 \overset{\mathrm{d}}{=} \mathrm{Poisson}(e^{X_{it}'\theta_0 + \alpha_{i0}})$, so that the likelihood for the $i$th individual is $L_{iT}(\theta, \alpha_i) = (\prod_{t=1}^{T} Y_{it}!)^{-1} e^{-\sum_{t=1}^{T} e^{X_{it}'\theta + \alpha_i}} e^{\sum_{t=1}^{T} Y_{it}(X_{it}'\theta + \alpha_i)}$. Hence, $\ell_{iT\alpha}(\theta, \alpha_i) = T^{-1}(-\sum_{t=1}^{T} e^{X_{it}'\theta + \alpha_i} + \sum_{t=1}^{T} Y_{it})$ and

$$\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \check{\theta}, \check{\alpha}] = T^{-1}(-\sum_{t=1}^{T} e^{X_{it}'\theta + \alpha_i} + \sum_{t=1}^{T} e^{X_{it}'\check{\theta} + \check{\alpha}}).$$

Consequently, the ZSE transformation and its inverse in the panel poisson model are

$$g_{iT\theta_0\theta}(\alpha_i) = \alpha_i + \log(\frac{\sum_{t=1}^{T} e^{X_{it}'\theta}}{\sum_{t=1}^{T} e^{X_{it}'\theta_0}}) \qquad \& \qquad h_{iT\theta_0\theta}(\phi) = \phi + \log(\frac{\sum_{t=1}^{T} e^{X_{it}'\theta_0}}{\sum_{t=1}^{T} e^{X_{it}'\theta}}).$$

It follows that $\tilde{L}_{iT}^0(\theta, \phi) = C_{iT}(\phi, \theta_0) \frac{e^{\sum_{t=1}^{T} Y_{it} X_{it}'\theta}}{(\sum_{t=1}^{T} e^{X_{it}'\theta}) \sum_{t=1}^{T} Y_{it}}$, where the constant $C_{iT}(\phi, \theta_0)$ does not depend upon $\theta$. Hence, modulo a factor that does not depend on $\theta$, we have that

$$\bar{L}_{iT}(\theta) = \frac{e^{\sum_{t=1}^{T} Y_{it} X_{it}'\theta}}{(\sum_{t=1}^{T} e^{X_{it}'\theta}) \sum_{t=1}^{T} Y_{it}} \propto \mathrm{pmf}_{\mathcal{Y}_{iT} | \mathcal{X}_{iT}, \sum_{t=1}^{T} Y_{it}; \theta}(\mathcal{Y}_{iT});$$

i.e., the ZSE transformed IL for individual $i$ is just the conditional density of $\mathcal{Y}_{iT}$ given $\mathcal{X}_{iT}$ and the statistic $\sum_{t=1}^{T} Y_{it}$, which is sufficient for $\alpha_i$. The MILE therefore coincides with the CMLE, which is the fixed effects MLE for the Poisson family (Lancaster, 2002, p. 650). □

**Example 9.3** (Panel logit)**.** Let $Y_{it} = \mathbb{1}(X_{it}'\theta_0 + \alpha_{i0} + U_{it} > 0)$, where $U_{i1}, \ldots, U_{iT} \big| \mathcal{X}_{iT}, \alpha_{i0} \overset{\mathrm{d}}{=} \text{LogisticIID}$. Since $\Pr(Y_{it} = y | \mathcal{X}_{iT}, \alpha_{i0}; \theta_0) = (\Lambda(X_{it}'\theta_0 + \alpha_{i0}))^y (1 - \Lambda(X_{it}'\theta_0 + \alpha_{i0}))^{1-y} \mathbb{1}_{\{0,1\}}(y)$, $y \in \mathbb{R}$, and the observations are independent across $t$, the likelihood for the $i$th individual is $L_{iT}(\theta, \alpha_i) = \prod_{t=1}^T (\Lambda(X_{it}'\theta + \alpha_i))^{Y_{it}} (1 - \Lambda(X_{it}'\theta + \alpha_i))^{1-Y_{it}}$. Consequently, $\ell_{iT\alpha}(\theta, \alpha_i) = T^{-1} \sum_{t=1}^T (Y_{it} - \Lambda(X_{it}'\theta + \alpha_i))$ and

$$\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \check{\theta}, \check{\alpha}] = T^{-1} \sum_{t=1}^T [\Lambda(X_{it}'\check{\theta} + \check{\alpha}) - \Lambda(X_{it}'\theta + \alpha_i)].$$

Using the MLE $\tilde{\theta}$ as the preliminary estimator, the inverse ZSE transformation $h_{iT\tilde{\theta}\theta}$ solves[11]

$$\sum_{t=1}^T [\Lambda(X_{it}'\tilde{\theta} + \phi) - \Lambda(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi))] = 0, \qquad \phi \in \mathbb{R}. \tag{9.2}$$

Unlike the previous examples, there is no closed form solution for $h_{iT\tilde{\theta}\theta}$. Consequently, there is also no closed form expression for the ZSE transformed IL $\bar{L}_{iT}(\theta) = \int_{\mathbb{R}} (\prod_{t=1}^T (\Lambda(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi)))^{Y_{it}} (1 - \Lambda(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi)))^{1-Y_{it}} \, d\phi$, which has to be obtained numerically as well.

In this example, we can use Corollary 6.1 to get some intuition behind the form of the ZSE transformed IL. Begin by observing that here $\ell_{iT\alpha\alpha}(\theta, \alpha_i) = -T^{-1} \sum_{t=1}^T \lambda(X_{it}'\theta + \alpha_i)$. Moreover, differentiating (9.2) with respect to $\phi$, we get that $\partial_\phi h_{iT\tilde{\theta}\theta}(\phi) = \sum_{t=1}^T \lambda(X_{it}'\tilde{\theta} + \phi) / \sum_{t=1}^T \lambda(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi))$. Hence, since $h_{iT\tilde{\theta}\theta}(\hat{\phi}_{iT\tilde{\theta}}(\theta)) = \hat{\alpha}_{iT}(\theta)$ by (D.14), from Corollary 6.1 we have that

$$\bar{L}_{iT}(\theta) = \sqrt{\frac{2\pi}{T}} L_{iT}(\theta, \hat{\alpha}_{iT}(\theta)) \Big(\frac{1}{T} \sum_{t=1}^T \lambda(X_{it}'\theta + \hat{\alpha}_{iT}(\theta))\Big)^{1/2} \frac{1}{T^{-1} \sum_{t=1}^T \lambda(X_{it}'\tilde{\theta} + \hat{\phi}_{iT\tilde{\theta}}(\theta))} (1 + O_{\mathrm{p}}(\frac{1}{T}))$$

for $\theta \in \Theta$. It is shown in Appendix H that $\hat{\phi}_{iT\tilde{\theta}}(\theta)$ is constant in $\theta$, in particular,

$$\hat{\phi}_{iT\tilde{\theta}}(\theta) = \hat{\alpha}_{iT}(\tilde{\theta}), \qquad \theta \in \Theta, \tag{9.3}$$

which can interpreted as meaning that strong unrelatedness holds in the panel logit model at the sample level as well. Consequently, the expression for $\bar{L}_{iT}(\theta)$ simplifies to

$$\bar{L}_{iT}(\theta) = M_{iT} L_{iT}(\theta, \hat{\alpha}_{iT}(\theta)) \Big(T^{-1} \sum_{t=1}^T \lambda(X_{it}'\theta + \hat{\alpha}_{iT}(\theta))\Big)^{1/2} (1 + O_{\mathrm{p}}(T^{-1})), \qquad \theta \in \Theta, \tag{9.4}$$

where $M_{iT} := \sqrt{2\pi/T} / T^{-1} \sum_{t=1}^T \lambda(X_{it}'\tilde{\theta} + \hat{\alpha}_{iT}(\tilde{\theta}))$ does not depend on $\theta$. As noted in Arellano (2003a, p. 443), the leading term in (9.4) is the modified profile likelihood for panel logit. It is also a saddlepoint approximation to the density of $\mathcal{Y}_{iT}$ given $\mathcal{X}_{iT}$ and the statistic $\sum_{t=1}^T Y_{it}$,

[11] In this example, a direct argument can be used to show that $\mathcal{D}(h_{iT\tilde{\theta}\theta}) = \mathbb{R}$ for each $\tilde{\theta}, \theta \in \Theta$, even though there is no closed form solution for $h_{iT\tilde{\theta}\theta}$. Indeed, given $\tilde{\theta}, \theta \in \Theta$, let $A(h) := \sum_{t=1}^T \Lambda(X_{it}'\theta + h)$ and $r_\phi := \sum_{t=1}^T \Lambda(X_{it}'\tilde{\theta} + \phi)$. The equation $A(h) = r_\phi$ can be solved for all $\phi \in \mathbb{R}$, because $h \mapsto A(h)$ is strictly increasing on $\mathbb{R}$ due to the fact $dA(h)/dh = \sum_{t=1}^T \lambda(X_{it}'\theta + h) > 0$.

which is sufficient for $\alpha_i$ (Levin, 1990, p. 278). This suggests that $\hat{\theta}$ is a good approximation of the CMLE. The approximation property of the MILE appears to hold even for small $T$ (cf. Table 1), although it is obtained using the fixed effects MLE (which is inconsistent as $n \to \infty$, $T$-fixed) as the preliminary estimator.

We end this example by explicitly demonstrating that the ZSE parameter is strongly unrelated to $\theta$ at the population level. Since $h_{iT\theta_0\theta}$ satisfies (9.2) when $\tilde{\theta}$ is replaced by $\theta_0$, it follows that $\phi^*_{iT}(\theta)$ satisfies

$$\begin{aligned}\sum_{t=1}^{T} \Lambda(X_{it}'\theta_0 + \phi^*_{iT}(\theta)) &= \sum_{t=1}^{T} \Lambda(X_{it}'\theta + h_{iT\theta_0\theta}(\phi^*_{iT}(\theta))) && (\theta \in \Theta)\\ &\overset{\text{(D.10)}}{=} \sum_{t=1}^{T} \Lambda(X_{it}'\theta + \alpha^*_{iT}(\theta))\\ &= \sum_{t=1}^{T} \Lambda(X_{it}'\theta_0 + \alpha_{i0}). && (\text{FOC of } \alpha^*_{iT}(\theta))\end{aligned}$$

Hence, $\phi^*_{iT}(\theta)$ is constant in $\theta$, with $\phi^*_{iT}(\theta) = \phi^*_{iT}(\theta_0) = \alpha_{i0}$ for each $\theta$. In other words, the ZSE parameter is strongly unrelated to $\theta$ at the population level (compare this with (9.3)). □

**Example 9.4** (Panel probit)**.** Let $Y_{it} = \mathbb{1}(X_{it}'\theta_0 + \alpha_{i0} + U_{it} > 0)$, where $U_{i1}, \ldots, U_{iT} \big| \mathcal{X}_{iT}, \alpha_{i0} \overset{\text{d}}{=} \text{NIID}(0,1)$. Since the observations are independent across $t$, the likelihood for the $i$th individual is $L_{iT}(\theta, \alpha_i) = \prod_{t=1}^{T} (\mathfrak{N}(X_{it}'\theta + \alpha_i))^{Y_{it}} (1 - \mathfrak{N}(X_{it}'\theta + \alpha_i))^{1-Y_{it}}$. Hence,

$$\mathbb{E}[\ell_{iT\alpha}(\theta, \alpha_i); \check{\theta}, \check{\alpha}] = T^{-1} \sum_{t=1}^{T} (\mathfrak{N}(X_{it}'\check{\theta} + \check{\alpha}) - \mathfrak{N}(X_{it}'\theta + \alpha_i)) \mathfrak{G}(X_{it}'\theta + \alpha_i),$$

where $\mathfrak{G} := \mathfrak{n}/\mathfrak{N}(1 - \mathfrak{N}) > 0$ is the probit weight-function. Therefore, using the fixed effects MLE as the preliminary estimator, the inverse ZSE transformation $h_{iT\tilde{\theta}\theta}$ solves[12]

$$\sum_{t=1}^{T} (\mathfrak{N}(X_{it}'\tilde{\theta} + \phi) - \mathfrak{N}(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi))) \mathfrak{G}(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi)) = 0, \qquad \phi \in \mathbb{R}. \tag{9.5}$$

There is no closed form solution for $h_{iT\tilde{\theta}\theta}$. As in Example 9.3, $h_{iT\tilde{\theta}\theta}$ and the ZSE transformed IL $\bar{L}_{iT}(\theta) = \int_{\mathbb{R}} (\prod_{t=1}^{T} (\mathfrak{N}(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi)))^{Y_{it}} (1 - \mathfrak{N}(X_{it}'\theta + h_{iT\tilde{\theta}\theta}(\phi)))^{1-Y_{it}} \, d\phi$ must be obtained numerically. A simple approximation of $\bar{L}_{iT}(\theta)$ as in (9.4) is not available in this example because $\theta \mapsto \hat{\phi}_{iT\tilde{\theta}}(\theta)$ is not constant for panel probit. □

[12] As with panel logit, a direct argument can be used to show that $\mathcal{D}(h_{iT\theta_0\theta}) = \mathbb{R}$ for each $\theta_0, \theta \in \Theta$. Let $\phi \in \mathbb{R}$ and note that a unique solution to the optimization problem $\max_{u \in \mathbb{R}} \mathbb{E}[\ell_{iT}(\theta, u); \theta_0, \phi]$ exists for each $\theta_0, \theta \in \Theta$ because $u \mapsto \mathbb{E}[\ell_{iT}(\theta, u); \theta_0, \phi]$ is strictly concave on $\mathbb{R}$, which follows from the strict concavity of the probit loglikelihood (Amemiya, 1985, Section 9.2.3) and the monotonicity of integrals. Hence, following the discussion after Lemma 4.1, $h_{iT\theta_0\theta} : \mathbb{R} \to \mathbb{R}$ exists for each $\theta_0, \theta \in \Theta$.

9.2. **Dynamic model.** In this section, we illustrate how our approach applies to a dynamic version of the Neyman-Scott model, namely, a stationary Gaussian AR(1) panel data model, without imposing time-independence on the response variables. The proofs of the results in this section are independent from the rest of the paper, and do not require the assumptions made in Appendix C.2–C.4. Using a direct approach, it is shown that the individual scores for the parameter of interest in the infeasible ZSE transformed IL are exactly unbiased, the information in the ZSE transformed IL is first-order unbiased, the MILE is consistent as $n, T_n \to \infty$, and its distribution is correctly centered as $n/T_n^3 \to 0$.

**Example 9.5** (Dynamic Neyman-Scott model)**.** Let $Y_{it} = \mu_0 Y_{i,t-1} + \alpha_{i0} + U_{it}$, $t = 1, \dots, T$, where $\mu_0 \in (-1, 1)$, $Y_{i0}$ is assumed to be observed (AB, p. 514), and

$$(U_{i1}, \dots, U_{iT}) \big| Y_{i0}, \alpha_{i0} \overset{\mathrm{d}}{=} \mathrm{NIID}(0, \sigma_0^2). \tag{9.6}$$

Here, $\mathcal{X}_{iT} = Y_{i0}$. Hence, the likelihood for the $i$th individual, conditional on $Y_{i0}$, is

$$L_{iT}(\mu, \sigma^2, \alpha_i) = (2\pi\sigma^2)^{-T/2} \exp(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} (Y_{it} - \mu Y_{i,t-1} - \alpha_i)^2), \tag{9.7}$$

which implies that $\ell_{iT\alpha}(\mu, \sigma^2, \alpha_i) = \frac{1}{T}[-\frac{T\alpha_i}{\sigma^2} - \frac{\mu}{\sigma^2} Y_{i0} + \frac{1}{\sigma^2} Y_{iT} + \frac{1-\mu}{\sigma^2} \sum_{t=1}^{T-1} Y_{it}]$. Hence,

$$\mathbb{E}[\ell_{iT\alpha}(\mu, \sigma^2, \alpha_i); \check{\mu}, \check{\sigma}^2, \check{\alpha}] = \frac{1}{T}[-\frac{T\alpha_i}{\sigma^2} - \frac{\mu}{\sigma^2} Y_{i0} + \frac{1}{\sigma^2} \beta_T + \frac{1-\mu}{\sigma^2} \sum_{s=1}^{T-1} \beta_s], \tag{9.8}$$

where $\beta_s := \int_{\mathbb{R}^T} y_{is} \, \mathrm{pdf}_{\mathrm{N}(\check{\mu} Y_{i0} + \check{\alpha}, \check{\sigma}^2)}(y_{i1}) \prod_{t=2}^{T} \mathrm{pdf}_{\mathrm{N}(\check{\mu} y_{i,t-1} + \check{\alpha}, \check{\sigma}^2)}(y_{it}) \, dy_{iT} \dots dy_{i1}$. It is shown in Appendix H that

$$\beta_s = \check{\alpha} \sum_{l=0}^{s-1} \check{\mu}^l + \check{\mu}^s Y_{i0}, \qquad s = 1, \dots, T. \tag{9.9}$$

Plugging (9.9) in (9.8), we get $\mathbb{E}[\ell_{iT\alpha}(\mu, \sigma^2, \alpha_i); \check{\mu}, \check{\sigma}^2, \check{\alpha}] = \frac{1}{T\sigma^2}[-T\alpha_i - a(\mu, \check{\mu}) Y_{i0} + \check{\alpha} c(\mu, \check{\mu})]$, where $a(\mu, \check{\mu}) := \mu - \check{\mu}^T - (1-\mu)\frac{\check{\mu} - \check{\mu}^T}{1-\check{\mu}}$, $c(\mu, \check{\mu}) := \frac{1-\check{\mu}^T}{1-\check{\mu}} + (1-\mu)\kappa(\check{\mu})$, and $\kappa(\check{\mu}) := \frac{1}{1-\check{\mu}}(T - \frac{1-\check{\mu}^T}{1-\check{\mu}})$. Thus, letting $\theta := (\mu, \sigma^2)$, the ZSE transformation and its inverse are

$$g_{iT\theta_0\theta}(\alpha_i) = (T\alpha_i + a(\mu, \mu_0) Y_{i0})/c(\mu, \mu_0) \quad \& \quad h_{iT\theta_0\theta}(\phi) = (\phi c(\mu, \mu_0) - a(\mu, \mu_0) Y_{i0})/T. \tag{9.10}$$

Using the expression for $h_{iT\theta_0\theta}$, it is shown in Appendix H that

$$\bar{L}_{iT}^0(\mu, \sigma^2) = \sqrt{T} (\sigma^2)^{-(T-1)/2} \exp(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} (\ddot{Y}_{it} - \mu \ddot{Y}_{i,t-1})^2)/c(\mu, \mu_0), \tag{9.11}$$

where $\ddot{Y}_{it} := Y_{it} - \bar{Y}_{i\cdot}$, $\ddot{Y}_{i,t-1} := Y_{i,t-1} - \bar{Y}_{i,-1}$, and $\bar{Y}_{i,-1} := \sum_{t=1}^{T} Y_{i,t-1}/T$. Replacing $\mu_0$ in (9.11) by the fixed effects MLE $\tilde{\mu} := \sum_{i=1}^{n} \sum_{t=1}^{T} \ddot{Y}_{it} \ddot{Y}_{i,t-1} / \sum_{i=1}^{n} \sum_{t=1}^{T} \ddot{Y}_{i,t-1}^2$, it follows that the feasible

ZSE transformed IL for the $i$th individual is given by

$$\bar{L}_{iT}(\mu,\sigma^2) = \sqrt{T}(\sigma^2)^{-(T-1)/2}\exp(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})^2)/c(\mu,\tilde{\mu}) \tag{9.12}$$

$$= L_{iT}^{\mathrm{p}}(\mu,\sigma^2)\sqrt{T}\sigma/c(\mu,\tilde{\mu}), \tag{9.13}$$

where (9.13) follows because $\hat{\alpha}_{iT}(\mu,\sigma^2) = \bar{Y}_{i\cdot} - \mu\bar{Y}_{i,-1}$ and the profile likelihood $L_{iT}^{\mathrm{p}}(\mu,\sigma^2) := L_{iT}(\mu,\sigma^2,\hat{\alpha}_{iT}(\mu,\sigma^2)) = (2\pi\sigma^2)^{-T/2}\exp(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})^2)$. Hence, our IL can be regarded as a "modified" version of the profile likelihood with correction factor $\sqrt{T}\sigma/c(\mu,\tilde{\mu})$.

The MILE $(\hat{\mu},\hat{\sigma}^2)$ solves $\nabla_{(\mu,\sigma^2)}\sum_{i=1}^{n}\log\bar{L}_{iT}(\hat{\mu},\hat{\sigma}^2) = 0$, i.e., $\hat{\mu}$ and $\hat{\sigma}^2$ jointly solve

$$\begin{aligned}\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{Y}_{i,t-1}(\ddot{Y}_{it}-\hat{\mu}\ddot{Y}_{i,t-1}) + \frac{n\kappa(\tilde{\mu})}{c(\hat{\mu},\tilde{\mu})} &= 0\\ \hat{\sigma}^2 - \frac{1}{n(T-1)}\sum_{i=1}^{n}\sum_{t=1}^{T}(\ddot{Y}_{it}-\hat{\mu}\ddot{Y}_{i,t-1})^2 &= 0.\end{aligned} \tag{9.14}$$

In the AR(1) example, as we now demonstrate, a direct approach can be used to show that the individual scores for $\mu$ in the infeasible IL are exactly unbiased compared to the individual scores for $\mu$ in the profile likelihood. This helps explain why $\hat{\mu}$ outperforms the fixed effects MLE $\tilde{\mu}$ markedly in the simulations. Begin by observing that since $c(\mu,\mu_0) > 0$ (cf. Footnote 50 in the supplementary material),

$$\bar{\ell}_{iT}^{0}(\mu,\sigma^2) = \frac{1}{T}[-\frac{(T-1)}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})^2 - \log c(\mu,\mu_0)]. \tag{9.15}$$

Hence, since $c(\mu_0,\mu_0) = T$,

$$\mathbb{E}_0\bar{\ell}_{iT\mu}^{0}(\mu_0,\sigma_0^2) = \frac{1}{T}[\frac{1}{\sigma_0^2}\sum_{t=1}^{T}\mathbb{E}_0(\ddot{Y}_{it}-\mu_0\ddot{Y}_{i,t-1})\ddot{Y}_{i,t-1} + \frac{\kappa(\mu_0)}{T}]. \tag{9.16}$$

It is shown in Appendix H that

$$\sum_{t=1}^{T}\mathbb{E}_0(\ddot{Y}_{it}-\mu_0\ddot{Y}_{i,t-1})\ddot{Y}_{i,t-1} = -\frac{\sigma_0^2}{1-\mu_0}(1-\frac{1}{T}\frac{1-\mu_0^T}{1-\mu_0}) = -\sigma_0^2\frac{\kappa(\mu_0)}{T}, \tag{9.17}$$

where the last equality follows from the definition of $\kappa$. Therefore, by (9.16),

$$\mathbb{E}_0\bar{\ell}_{iT\mu}^{0}(\mu_0,\sigma_0^2) = 0 \iff \mathbb{E}_0[\frac{1}{\sigma_0^2}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu_0\ddot{Y}_{i,t-1})\ddot{Y}_{i,t-1} + \frac{\kappa(\mu_0)}{T}] = 0, \tag{9.18}$$

i.e., the IL score for individual $i$ is unbiased for zero. In contrast, it can be shown that the bias of the profile likelihood score $\mathbb{E}_0\ell_{iT\mu}^{\mathrm{p}}(\mu_0,\sigma_0^2) = O(1/T)$ as $T\to\infty$. It follows that the individual scores for $\mu$ in the ZSE transformed IL are exactly unbiased compared to the individual scores for $\mu$ in the profile likelihood.

A direct approach can also be used to demonstrate that, for each individual, the information for $\mu$ in the ZSE transformed IL is first-order unbiased in the sense that the information equality for $\mu$ holds up to an error of order $T^{-2}$, i.e., as shown in Appendix H,

$$T\mathbb{E}_0[\overline{\ell}^0_{iT\mu}(\mu_0,\sigma_0^2)]^2+\mathbb{E}_0\overline{\ell}^0_{iT\mu\mu}(\mu_0,\sigma_0^2)=O(T^{-2}). \tag{9.19}$$

The asymptotic behavior of $\hat{\mu}$ and $\hat{\sigma}^2$ as $n,T\to\infty$ can also be obtained from first principles. To demonstrate this, we focus on $\hat{\mu}$. Let $\hat{a}:=(nT)^{-1}\sum_{i=1}^n\sum_{t=1}^T\ddot{Y}_{i,t-1}^2$, $\hat{v}(\mu):=(n(T-1))^{-1}\sum_{i=1}^n\sum_{t=1}^T(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})^2$, $b_T(\mu_1,\mu_2):=(\frac{1-\mu_2^T}{1-\mu_1}+T-\frac{1-\mu_2^T}{1-\mu_2})^{-1}(T-\frac{1-\mu_2^T}{1-\mu_2})$, and $\hat{\gamma}(\hat{\mu},\tilde{\mu}):=\hat{v}(\hat{\mu})b_T(\hat{\mu},\tilde{\mu})/\hat{a}$. The following result shows that $\hat{\mu}$ is consistent for $\mu_0$ as $n,T_n\to\infty$, by showing that $\hat{\mu}$ is a fixed point of a simple function.

**Lemma 9.1.** *Let $\mathbb{E}\alpha_{i0}^2+\mathbb{E}Y_{i0}^2<\infty$. Then, $\hat{\mu}$ solves $\hat{\mu}=[(1+\tilde{\mu})-\sqrt{(1-\tilde{\mu})^2-4T^{-1}\hat{\gamma}(\hat{\mu},\tilde{\mu})}]/2$, and is a consistent estimator of $\mu_0$ as $n,T_n\to\infty$.*

Since $\operatorname{plim}_{n\to\infty}\hat{\gamma}(\hat{\mu},\tilde{\mu})/T_n=0$ by Lemma H.1, the square-root term is well defined w.p.a.1. It is evident from Lemma 9.1 that consistency of $\hat{\mu}$ holds irrespective of whether $T_n\to\infty$ faster or slower than $n\to\infty$. Similarly, consistency of $\hat{\sigma}^2$ can be shown from (H.25) and Lemma H.3. Next, we describe the distribution of $\hat{\mu}$ as $n,T_n\to\infty$. In what follows, keep in mind that, in the AR(1) model, $\alpha^*_{iT}(\mu,\sigma^2)=[-a(\mu,\mu_0)Y_{i0}+\alpha_{i0}c(\mu,\mu_0)]/T=:\alpha^*_{iT}(\mu)$.

**Lemma 9.2.** *Let $\mathbb{E}\alpha_{i0}^2+\mathbb{E}Y_{i0}^2<\infty$. Then,*

$$\begin{aligned}&(\hat{a}-O_{\mathrm{p}}(T_n^{-1}))\sqrt{nT_n}(\hat{\mu}-\mu_0)\\&\qquad=\frac{1}{\sqrt{nT_n}}\sum_{i=1}^n\sum_{t=1}^{T_n}U_{it}(Y_{i,t-1}+\partial_\mu\alpha^*_{iT}(\mu_0))+O_{\mathrm{p}}(\frac{1}{\sqrt{T_n}})+O_{\mathrm{p}}(\sqrt{\frac{n}{T_n^3}}).\end{aligned} \tag{9.20}$$

*Furthermore, if $\mathbb{E}\alpha_{i0}^4+\mathbb{E}Y_{i0}^4<\infty$, and $n/T_n^3\to0$ as $n,T_n\to\infty$, then*

$$\sqrt{nT_n}(\hat{\mu}-\mu_0)\xrightarrow{\mathrm{d}}\mathrm{N}(0,1-\mu_0^2). \tag{9.21}$$

Existence of the fourth moments $\mathbb{E}\alpha_{i0}^4,\mathbb{E}Y_{i0}^4$ is used to show that Lindeberg's condition holds for $(nT_n)^{-1/2}\sum_{i=1}^n\sum_{t=1}^{T_n}U_{it}(Y_{i,t-1}+\partial_\mu\alpha^*_{iT}(\mu_0))$. Lemma 9.2 reveals that the asymptotic variance of $\sqrt{nT_n}(\hat{\mu}-\mu_0)$ equals that of the bias corrected estimator of Hahn and Kuersteiner (2002, Section 3, p. 1645), which coincides with the asymptotic variance of the fixed effects MLE (Hahn and Kuersteiner, Theorem 1). Unlike Hahn and Kuersteiner, who use $\lim_{n\to\infty}n/T_n\in(0,\infty)$ to prove their result, cf. their Condition 4, we only need the weaker condition $\lim_{n\to\infty}n/T_n^3=0$. Although Hahn and Kuersteiner derive their bias correction for AR(1) models under a much weaker assumption on the model errors (compare (9.6) with Condition 1(i) on p. 1641 of their paper), their Gaussian special case is the same as Example 9.5 (compare (9.6) and the assumption $\mathbb{E}\alpha_{i0}^2+\mathbb{E}Y_{i0}^2<\infty$ with their Condition 4).

**Remark 9.1.** (i) It is instructive to examine the moment conditions solved by the MILE in this example. From (9.15), it is clear that

$$\begin{aligned}\bar{\ell}^0_{iT\mu}(\mu,\sigma^2) &= \frac{1}{T}[\frac{1}{\sigma^2}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})\ddot{Y}_{i,t-1}+\frac{\kappa(\mu_0)}{c(\mu,\mu_0)}] \\ \bar{\ell}^0_{iT\sigma^2}(\mu,\sigma^2) &= -\frac{(T-1)}{T}\frac{1}{2\sigma^4}[\sigma^2-\frac{1}{T-1}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})^2].\end{aligned} \tag{9.22}$$

Note that $\bar{\ell}^0_{iT\mu}(\mu,\sigma^2)$ depends upon $\mu_0$, which is estimated by a preliminary estimator. To allow for this, let $\gamma_0 := \mu_0$ and assume that $\gamma_0$ is identified via the single moment condition $\mathbb{E}_0 m(\mathcal{Y}_{iT0},\gamma_0)=0$, where $\mathcal{Y}_{iT0} := (Y_{i0},Y_{i1},\ldots,Y_{iT})$; e.g., if $\gamma_0$ is estimated by the fixed effects MLE, then $m(\mathcal{Y}_{iT0},\gamma) := \sum_{t=1}^{T}(\ddot{Y}_{it}-\gamma\ddot{Y}_{i,t-1})\ddot{Y}_{i,t-1}$. Comparing (9.14) and (9.22), the MILE of $(\mu_0,\sigma_0^2)$ solves the exactly-identified system of moment conditions $\mathbb{E}_0\rho(\mathcal{Y}_{iT0},\mu_0,\sigma_0^2,\gamma_0)=0_{3\times1}$, where $\rho(\mathcal{Y}_{iT0},\mu,\sigma^2,\gamma) := (\rho_1(\mathcal{Y}_{iT0},\mu,\sigma^2,\gamma),\rho_2(\mathcal{Y}_{iT0},\mu,\sigma^2),\rho_3(\mathcal{Y}_{iT0},\gamma))_{3\times1}$ and

$$\begin{aligned}\rho_1(\mathcal{Y}_{iT0},\mu,\sigma^2,\gamma) &:= \frac{1}{\sigma^2}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})\ddot{Y}_{i,t-1}+\frac{\kappa(\gamma)}{c(\mu,\gamma)} \\ \rho_2(\mathcal{Y}_{iT0},\mu,\sigma^2) &:= \sigma^2-\frac{1}{T-1}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})^2 \\ \rho_3(\mathcal{Y}_{iT0},\gamma) &:= m(\mathcal{Y}_{iT0},\gamma).\end{aligned}$$

Hence, the MILE $(\hat{\mu},\hat{\sigma}^2)$ is obtained by solving $n^{-1}\sum_{i=1}^{n}\rho(\mathcal{Y}_{iT0},\hat{\mu},\hat{\sigma}^2,\hat{\gamma})=0_{3\times1}$. This suggests that the MILE of $(\mu_0,\sigma_0^2)$ may be consistent and asymptotically normal as $n\to\infty$, $T$-fixed, provided the preliminary estimator $\hat{\gamma}$ is also consistent and asymptotically normal as $n\to\infty$, $T$-fixed. This rules out $\hat{\gamma}$ being the fixed effects MLE, because it well known that the fixed effects MLE is inconsistent (as $n\to\infty$, $T$-fixed) in this example (Nickell, 1981). However, if, say, $\hat{\gamma}$ is the instrumental-variables (IV) estimator of Anderson and Hsiao (1981), obtained by letting $m(\mathcal{Y}_{iT0},\gamma) := \sum_{t=2}^{T}(\Delta Y_{it}-\gamma\Delta Y_{i,t-1})Y_{i,t-2}$, where $\Delta Y_{it} := Y_{it}-Y_{i,t-1}$, then $\hat{\gamma}$ is consistent as $n\to\infty$, $T$-fixed, and GMM theory can be used to show consistency and obtain the joint distribution of $n^{1/2}(\hat{\mu}-\mu_0)$ and $n^{1/2}(\hat{\sigma}^2-\sigma_0^2)$ as $n\to\infty$, $T$-fixed.

(ii) The MILE $\hat{\mu}$ differs from the estimator in Lancaster (2002, Equation 3.20) because the MILE solves the FOC $\sigma^{-2}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{Y}_{i,t-1}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})+n\kappa(\tilde{\mu})/c(\mu,\tilde{\mu})=0$, whereas Lancaster's estimator solves the FOC $\sigma^{-2}\sum_{i=1}^{n}\sum_{t=1}^{T}\ddot{Y}_{i,t-1}(\ddot{Y}_{it}-\mu\ddot{Y}_{i,t-1})+n\kappa(\mu)/T=0$. Indeed, Lancaster's estimator of $\mu$ is just the GMM estimator based on the moment condition (9.18). The limiting ($n\to\infty$, $T$-fixed) distribution of Lancaster's estimator of $(\mu,\sigma^2)$, which is not given in his paper, can therefore be easily obtained by applying GMM theory to (9.18) and the moment condition for $\sigma^2$, i.e., $\mathbb{E}_0[\sigma_0^2-(T-1)^{-1}\sum_{t=1}^{T}(\ddot{Y}_{it}-\mu_0\ddot{Y}_{i,t-1})^2]=0$. □

## 10. SIMULATION STUDY

We now examine the small sample behavior of the MILE ($\hat{\theta}$), and its iterated versions $\hat{\theta}(1)$ and $\hat{\theta}(\infty)$, in the logit, probit, and AR(1) models. The results in this section are based on 1000 Monte Carlo replications, and we use $\pi_i := 1$ and $(\mathfrak{a}, \mathfrak{b}) = \mathbb{R}$ to compute the ZSE transformed IL. To ensure a fair comparison between estimators based on the three different integrated likelihoods available in the literature — namely, Lancaster's IOT based IL, AB's weighted IL, and our ZSE transformed IL — the performance of the MILE is compared with Lancaster's estimator, two versions of $\hat{\theta}_{\text{AB}}$ ("observed" and "expected"), the fixed effects MLE, and some other estimators, for $n = 100, 500$ and $T = 5, 10, 20$. Results for $n = 100$ (Tables 1–6) are in the paper. Results for $n = 500$ (Tables 7–12), and all figures relating to the simulation results, can be found in Appendix A of the supplementary material.

In the context of AB, "observed" refers to the IL of AB using the "robust" weight-function, defined in their Equation 12, with expectations replaced by time-averages, whereas "expected" is the IL of AB where their robust weight-function is estimated using expected quantities with the true $\theta_0$ replaced by the MLE $\tilde{\theta}$ and the true $\alpha_{i0}$ replaced by $\hat{\alpha}_{iT}(\tilde{\theta})$. Note that, in their simulations, AB report $\hat{\theta}_{\text{AB}}$(observed) and $\hat{\theta}_{\text{AB}}$(infeasible), the latter being the oracle estimator based on expected quantities and the true values of $(\theta_0, \alpha_{i0})$; cf. the discussion on p. 519 of their paper.

10.1. **Designs.** The following designs are implemented in our simulation study.

10.1.1. *Logit.* We use the design in AB (Section 7.1), so that we can compare the MILE with their estimator. In other words, $Y_{it} := \mathbb{1}(X_{it}\theta_0 + \alpha_{i0} + U_{it} > 0)$, where $U_{i1}, \ldots, U_{iT} \overset{\text{d}}{=} \text{LogisticIID}$ and independent of $(\mathcal{X}_{iT}, \alpha_{i0})$, the regressors $\mathcal{X}_{iT} \overset{\text{d}}{=} \text{NIID}(0, 1)$, $\alpha_{i0} \overset{\text{d}}{=} \text{N}(\bar{X}_{i\cdot}, 1)$, and $\theta_0 = 1$. The MILE is compared with the fixed effects MLE, which is inconsistent as $n \to \infty$, $T$-fixed (Chamberlain, 1980, p. 228); the CMLE (obtained using the `clogit` function in the `survival` package), which is the benchmark for the panel logit model because it is consistent as $n \to \infty$, $T$-fixed (Andersen, 1970, Section 4); the IOT based estimator of Lancaster (2000, Section 6.6); and $\hat{\theta}_{\text{AB}}$(observed) and $\hat{\theta}_{\text{AB}}$(expected), implemented using Equations 36 and 37, respectively, of AB. Unlike AB, we also report results for $n = 500$. The empirical coverage of the LR confidence region constructed using the MILE and its iterated versions is compared with that based on $\hat{\theta}_{\text{AB}}$(observed, expected), Lancaster's estimator, and the MLE.

10.1.2. *Probit.* Here, $Y_{it} := \mathbb{1}(X_{it}\theta_0 + \alpha_{i0} + U_{it} > 0)$, where $U_{i1}, \ldots, U_{iT} \overset{\text{d}}{=} \text{NIID}(0, 1)$ and are independent of $(\mathcal{X}_{iT}, \alpha_{i0})$, the regressors $\mathcal{X}_{iT} \overset{\text{d}}{=} \text{NIID}(0, 1)$, $\alpha_{i0} \overset{\text{d}}{=} \text{N}(\bar{X}_{i\cdot}, 1)$, and $\theta_0 = 1$. The MILE is compared with the fixed effects MLE, which is inconsistent in the $n \to \infty$, $T$-fixed setting, $\hat{\theta}_{\text{AB}}$(observed, expected) implemented using Equation 12 of AB, and the IOT based estimator of Lancaster (2000, Section 6.7). Note that AB do not report simulation results

for panel probit in their paper. Also reported are the empirical coverage probabilities of the LR confidence regions based on the MILE and its iterated versions, $\hat{\theta}_{\text{AB}}$(observed, expected), Lancaster's estimator, and the MLE.

10.1.3. *AR*(1). Again, we use the design in AB (Section 7.2) so that we can compare the MILE with their estimator; i.e., $Y_{it} := \mu_0 Y_{i,t-1} + \alpha_{i0} + U_{it}$, where $U_{i1}, \ldots, U_{iT} \big| Y_{i0}, \alpha_{i0} \overset{\text{d}}{=} \text{NIID}(0, \sigma_0^2)$. The initial condition is drawn from the stationary distribution of $Y_{it}$, i.e., $Y_{i0} \overset{\text{d}}{=} \text{N}(\alpha_{i0}/(1-\mu_0), \sigma_0^2/(1-\mu_0^2))$, $\alpha_{i0} \overset{\text{d}}{=} \text{N}(0,1)$, $\mu_0 = 0.5$, and $\sigma_0^2 = 1$. The variance of the error term $\sigma_0^2$ is treated as known; the objective is to estimate $\mu_0$. The MILE $\hat{\mu}$ is compared with the fixed effects MLE $\tilde{\mu}$, which is inconsistent as $n \to \infty$ and $T$ is fixed; the IV estimator of Anderson and Hsiao, the GMM estimator of Arellano and Bond (1991) based on the sequential moment conditions $\mathbb{E}Y_{i,t-k}(\Delta Y_{it} - \mu \Delta Y_{i,t-1}) = 0$, $t = 2, \ldots, T$, $k = 2, \ldots, t$; the GMM estimator obtained by pooling these $T(T-1)/2$ moment conditions (the IV and GMM estimators are consistent as $n \to \infty$, $T$-fixed); the IOT based estimator of Lancaster (2000, Section 6.5); $\hat{\mu}_{\text{AB}}$(observed) implemented using Equation 14 of AB, $\hat{\mu}_{\text{AB}}$(expected) implemented using Equation 31 of AB, and the iterated version of $\hat{\mu}_{\text{AB}}$(expected).

The weight-function for $\hat{\theta}_{\text{AB}}$(observed) requires an estimator of the long-run ($T \to \infty$) variance of $\ell_{iT\alpha}$. Since the observations in the AR(1) model are serially correlated, this was obtained using the HAC estimator described in Equations 8 and 9 of Arellano and Hahn (2016) constructed with the Bartlett kernel (note the typo in the expression for the Bartlett kernel given on p. 257 of their paper) and bandwidth = 2. We set the bandwidth parameter, referred to as the "degree of trimming" in Section 7.2 of AB, equal to 2 because, as reported by AB in their Table II (p. 523), that value produced the smallest MSE for $\hat{\mu}_{\text{AB}}$(observed).

In addition, we also report the empirical coverage probabilities of LR confidence regions based on the MILE and its iterated versions, $\hat{\mu}_{\text{AB}}$(observed, expected), the iterated version of $\hat{\mu}_{\text{AB}}$(expected), Lancaster's estimator, and the MLE.

10.2. **Results.** The main findings of our simulation study, which follow our theoretical results fairly closely, can be summarized as follows (a detailed discussion is in Appendix A.2):

(i) The bias of the MILE, Lancaster's estimator, $\hat{\theta}_{\text{AB}}$(observed, expected), and the fixed effects MLE, is driven primarily by $T$, i.e., in each model, the biases of these estimators for $T = 5, 10, 20$ are roughly the same for $n = 100$ and $n = 500$.

(ii) In the logit and probit designs, the MILE outperforms $\hat{\theta}_{\text{AB}}$(observed, expected) and the fixed effects MLE, and narrowly beats the MSE of Lancaster's estimator, when $T = 5$. Indeed, for $T = 5$, there appears to be little difference between the distribution of $\hat{\theta}_{\text{AB}}$(observed) and the fixed effects MLE. Although the distribution of $\hat{\theta}_{\text{AB}}$(observed) approaches the distribution of the MILE as $T$ grows, its bias is still large compared to the MILE when $T = 20$. $\hat{\theta}_{\text{AB}}$(expected) exhibits significantly less bias than $\hat{\theta}_{\text{AB}}$(observed)

for small $T$, although its bias is much larger than that of the MILE in the probit design even when $T = 20$. The performance of Lancaster's estimator is quite comparable to the MILE and better than $\hat{\theta}_{\text{AB}}$(expected). When $T$ is small, the coverage probability of the MILE-based LR confidence region is much closer to its nominal value (95%) than those based on $\hat{\theta}_{\text{AB}}$(observed, expected), Lancaster's estimator, and the fixed effects MLE. A comparison of the simulation results for the logit and probit designs suggests that the MILE and Lancaster's estimator work well for both with stable performance in terms of bias. In contrast, $\hat{\theta}_{\text{AB}}$(observed, expected) appears to work better for logit than for probit, with higher bias for the probit design. The fixed effects MLE performs poorly in both designs, though its bias is worse for probit.

(iii) In the AR(1) design, the MILE or one of its iterated versions are the best in terms of the MSE for each $T$. The MILE exhibits some (upwards) bias when $T = 5$, but the bias decreases rapidly as $T$ grows, and for $T = 20$ it is almost perfectly centered at the truth. In stark contrast, $\hat{\mu}_{\text{AB}}$(observed) and the fixed effects MLE are substantially downwards biased when $T = 5$. $\hat{\mu}_{\text{AB}}$(expected) has significantly less bias than $\hat{\mu}_{\text{AB}}$(observed), and also less bias than the MILE (but not its iterated versions) when $T = 5$. Although the magnitude of their bias decreases as $T$ grows, $\hat{\mu}_{\text{AB}}$(observed) and the fixed effects MLE are still substantially downwards biased even when $T = 20$. Lancaster's estimator, which does not require any preliminary estimator, performs really well, e.g., it has the least bias for $T = 5$. In each period, the iterated version of $\hat{\mu}_{\text{AB}}$(expected) exhibits significantly less bias than $\hat{\mu}_{\text{AB}}$(expected) itself. Indeed, when $T = 5$, it is only narrowly beaten by MILE($\infty$) in terms of the MSE. For $T = 10, 20$, the bias and variance of the MILE and its iterated versions, the iterated version of $\hat{\mu}_{\text{AB}}$(expected), and Lancaster's estimator, are virtually identical. The coverage probability of the LR confidence regions based on MILE(1) and Lancaster's estimator are much closer to the nominal value than their competitors. This makes sense because MILE(1) and Lancaster's estimator behave similarly in terms of their bias and variance. The same reason explains the low coverage probabilities of the confidence regions based on $\hat{\theta}_{\text{AB}}$(observed) and the fixed effects MLE.

As noted in Section 8.3, the finding that the MILE-based LR confidence region has better empirical coverage than the $\hat{\theta}_{\text{AB}}$-based LR can be attributed to the fact that the ZSE transformed IL is first-order information unbiased, whereas the IL of AB is not. This also explains why, in the AR(1) design, even though the iterated version of $\hat{\mu}_{\text{AB}}$(expected) beats the iterated version of the MILE narrowly in terms of the MSE when $T = 5$, the coverage probability of its LR confidence regions, unlike that of the iterated MILE, is much below the nominal level. As expected, cf. the discussion following (8.5), the coverage of the profile likelihood based confidence region is poor for each $n, T$.

Let $H_{nT} := n^{-1}\sum_{i=1}^{n}\mathbb{E}_0\nabla^2_{\theta\theta}\bar{\ell}_{iT}(\theta_0)$ denote the expected hessian of the ZSE transformed IL, and $H_{nT}(\text{AB}) := n^{-1}\sum_{i=1}^{n}\mathbb{E}_0\nabla^2_{\theta\theta}\bar{\ell}^{\text{AB}}_{iT}(\theta_0)$ the expected hessian matrix of the IL of AB. One reason why the MILE can have smaller finite sample variance than $\hat{\theta}_{\text{AB}}$, at least when $T$ is small, follows from the results in SST, who show that, under certain conditions, as $n, T \to \infty$,

$$\begin{aligned}\text{var}_0\sqrt{nT}(\hat{\theta}-\theta_0) &= -H_{nT}^{-1} + O_{\text{p}}(\frac{n}{T^3}) + O_{\text{p}}(\frac{1}{T^2}) \\ \text{var}_0\sqrt{nT}(\hat{\theta}_{\text{AB}}-\theta_0) &= -H_{nT}^{-1}(\text{AB}) + O_{\text{p}}(\frac{n}{T^3}) + O_{\text{p}}(\frac{1}{T}),\end{aligned} \tag{10.1}$$

where, abusing notation, $\text{var}_0$ is now variance with respect to $\prod_{i=1}^{n} f_{\mathrm{y}_{iT}|\mathrm{x}_{iT},\alpha_{i0};\theta_0}$. The presence of the $O_{\text{p}}(T^{-2})$ term in the expression for $\text{var}_0\sqrt{nT}(\hat{\theta}-\theta_0)$, which is a consequence of the fact that the ZSE transformed IL is first-order information unbiased, thus provides mathematical justification behind why the MILE can have smaller variance than $\hat{\theta}_{\text{AB}}$ in finite samples when $T$ is small. However, our simulations reveal that the difference between the finite sample variance of $\hat{\theta}_{\text{AB}}$ and the MILE is more pronounced for $\hat{\theta}_{\text{AB}}$(observed) than $\hat{\theta}_{\text{AB}}$(expected). Moreover, the finite sample variance of the MILE is smaller than that of $\hat{\theta}_{\text{AB}}$(observed) and $\hat{\theta}_{\text{AB}}$(expected) in the logit and probit designs, but not in the AR(1) design. This suggests that (10.1), which is an asymptotic result, should be used with caution (as with all asymptotic results) to rank the finite sample variances of the MILE and competing estimators. Indeed, (10.1) reveals that the variance of the MILE and $\hat{\theta}_{\text{AB}}$ depend on the hessians and the constants in the $O_{\text{p}}$ terms. Therefore, unless these hessians and constants are comparable across estimators, the ability of (10.1) to rank finite sample variances can be limited.

Another reason why the MILE performs better than $\hat{\theta}_{\text{AB}}$(observed) may be due to the manner in which the AB weight-functions are constructed. The weight-functions used by $\hat{\theta}_{\text{AB}}$(observed) are obtained by replacing expectations with large-$T$ consistent sample averages (AB, Section 3.1). In other words, $\hat{\theta}_{\text{AB}}$(observed) uses individual-specific time-series when forming the weight-function so that, in finite samples, each weight-function is based on relatively little data (recall that we are dealing with short panels here). More precisely, for large-$T$ consistent estimators we expect the difference between the estimator and estimand to be $O_{\text{p}}(T^{-1/2})$. In contrast, the inverse ZSE transformation used to obtain the MILE depends on the entire sample because we replace $\theta_0$ by the fixed effects MLE when constructing the inverse ZSE transformation. However, the fixed effects MLE uses all of the data so that it tends to be more accurate than individual-specific estimators. Indeed, as $n, T \to \infty$, $\tilde{\theta}-\theta_0 \overset{\text{Ass. C.6(ii)}}{=} O_{\text{p}}((nT)^{-1/2}) + O_{\text{p}}(T^{-1}) \overset{(\frac{n}{T}\not\to 0)}{=} O_{\text{p}}(T^{-1})$. Thus, the estimator used to construct the ZSE transformed IL is more accurate than the estimator used by AB to construct their IL, which also helps explain why the bias of the MILE is smaller than the bias of $\hat{\theta}_{\text{AB}}$(observed). This also explains why $\hat{\theta}_{\text{AB}}$(expected) is less biased than $\hat{\theta}_{\text{AB}}$(observed), which agrees with the results in Schumann (2020).

## 11. CONCLUSION

We have demonstrated a new integrated likelihood based approach for estimating panel data models when the fixed effects enter the model nonlinearly. Unlike existing integrated likelihoods in the literature, the one we propose appears to be closer to the target likelihood because it reduces score and information bias simultaneously. Reduction in information bias is related to better performance of the MILE and the likelihood ratio statistic in panels of small durations. Results from a simulation study suggest that, in both static and dynamic models, our methodology can work very well even in moderately sized panels of short duration. One issue not addressed in this paper is that of estimating marginal effects. Although it is possible to estimate marginal effects using an integrated likelihood approach, addressing this complex issue requires a separate treatment. Research on this topic is in progress, and will be reported in another paper.

## ACKNOWLEDGEMENTS

We thank the editor Serena Ng, an associate editor, and three anonymous referees, for comments that greatly improved this paper. We also thank Antonio Cosma, Geert Dhaene, Arnaud Dupuy, Bernd Fitzenberger, Dennis Kristensen, Taisuke Otsu, Martin Weidner, and seminar participants at Aarhus, Bonn, Cologne, CORE, Dortmund, Humboldt, KU-Leuven, LSE, Luxembourg, Mannheim, UCL, the 2nd joint conference of the Belgian, Royal Spanish and Luxembourg Mathematical Societies in Rioja, and the 2016 European Meeting of the Econometric Society in Geneva, for helpful suggestions and conversations. The simulation experiments reported in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014, `http://hpc.uni.lu`).

TABLE 1. Simulation results for the logit model ($n = 100$).

| $T$ | Estimator | Mean Bias | STD | MSE | MAE | Median Bias |
|---|---|---|---|---|---|---|
| 5 | MILE | 0.0885 | 0.1435 | 0.0284 | 0.1336 | 0.0850 |
| | MILE(1) | 0.0885 | 0.1436 | 0.0285 | 0.1337 | 0.0850 |
| | MILE($\infty$) | 0.0885 | 0.1437 | 0.0285 | 0.1337 | 0.0850 |
| | AB(observed) | 0.2739 | 0.1879 | 0.1103 | 0.2824 | 0.2648 |
| | AB(expected) | 0.1418 | 0.1535 | 0.0437 | 0.1685 | 0.1392 |
| | MLE | 0.3271 | 0.1881 | 0.1423 | 0.3313 | 0.3216 |
| | Lancaster | 0.1002 | 0.1455 | 0.0312 | 0.1403 | 0.0974 |
| | CMLE | 0.0166 | 0.1346 | 0.0184 | 0.1081 | 0.0135 |
| 10 | MILE | 0.0175 | 0.0961 | 0.0095 | 0.0778 | 0.0151 |
| | MILE(1) | 0.0175 | 0.0961 | 0.0095 | 0.0777 | 0.0151 |
| | MILE($\infty$) | 0.0174 | 0.0960 | 0.0095 | 0.0777 | 0.0151 |
| | AB(observed) | 0.0577 | 0.1019 | 0.0137 | 0.0928 | 0.0541 |
| | AB(expected) | 0.0327 | 0.0981 | 0.0107 | 0.0821 | 0.0302 |
| | MLE | 0.1310 | 0.1102 | 0.0293 | 0.1420 | 0.1282 |
| | Lancaster | 0.0210 | 0.0966 | 0.0098 | 0.0787 | 0.0185 |
| | CMLE | 0.0014 | 0.0948 | 0.0090 | 0.0758 | −0.0008 |
| 20 | MILE | 0.0060 | 0.0636 | 0.0041 | 0.0505 | 0.0043 |
| | MILE(1) | 0.0060 | 0.0636 | 0.0041 | 0.0505 | 0.0043 |
| | MILE($\infty$) | 0.0060 | 0.0636 | 0.0041 | 0.0505 | 0.0043 |
| | AB(observed) | 0.0167 | 0.0646 | 0.0045 | 0.0526 | 0.0150 |
| | AB(expected) | 0.0109 | 0.0640 | 0.0042 | 0.0512 | 0.0091 |
| | MLE | 0.0623 | 0.0681 | 0.0085 | 0.0749 | 0.0604 |
| | Lancaster | 0.0075 | 0.0638 | 0.0041 | 0.0507 | 0.0056 |
| | CMLE | 0.0026 | 0.0635 | 0.0040 | 0.0503 | 0.0008 |

TABLE 2. Simulation results for the probit model ($n = 100$).

| $T$ | Estimator | Mean Bias | STD | MSE | MAE | Median Bias |
|---|---|---|---|---|---|---|
| 5 | MILE | 0.0835 | 0.1119 | 0.0195 | 0.1107 | 0.0793 |
| | MILE(1) | 0.0931 | 0.1137 | 0.0216 | 0.1171 | 0.0874 |
| | MILE($\infty$) | 0.0928 | 0.1136 | 0.0215 | 0.1169 | 0.0871 |
| | AB(observed) | 0.4303 | 0.2090 | 0.2288 | 0.4308 | 0.4096 |
| | AB(expected) | 0.1819 | 0.1283 | 0.0495 | 0.1885 | 0.1738 |
| | MLE | 0.3968 | 0.1668 | 0.1852 | 0.3970 | 0.3842 |
| | Lancaster | 0.1089 | 0.1146 | 0.0250 | 0.1276 | 0.1031 |
| 10 | MILE | 0.0199 | 0.0703 | 0.0053 | 0.0570 | 0.0159 |
| | MILE(1) | 0.0223 | 0.0704 | 0.0055 | 0.0576 | 0.0185 |
| | MILE($\infty$) | 0.0223 | 0.0704 | 0.0055 | 0.0576 | 0.0185 |
| | AB(observed) | 0.0924 | 0.0802 | 0.0150 | 0.1003 | 0.0875 |
| | AB(expected) | 0.0671 | 0.0757 | 0.0102 | 0.0807 | 0.0622 |
| | MLE | 0.1633 | 0.0860 | 0.0341 | 0.1646 | 0.1582 |
| | Lancaster | 0.0275 | 0.0707 | 0.0058 | 0.0592 | 0.0236 |
| 20 | MILE | 0.0063 | 0.0488 | 0.0024 | 0.0393 | 0.0037 |
| | MILE(1) | 0.0069 | 0.0488 | 0.0024 | 0.0394 | 0.0042 |
| | MILE($\infty$) | 0.0069 | 0.0488 | 0.0024 | 0.0394 | 0.0042 |
| | AB(observed) | 0.0248 | 0.0505 | 0.0032 | 0.0443 | 0.0219 |
| | AB(expected) | 0.0269 | 0.0504 | 0.0033 | 0.0451 | 0.0238 |
| | MLE | 0.0743 | 0.0538 | 0.0084 | 0.0778 | 0.0709 |
| | Lancaster | 0.0085 | 0.0489 | 0.0025 | 0.0396 | 0.0057 |

TABLE 3. Simulation results for the AR(1) model ($n = 100$).

| $T$ | Estimator | Mean Bias | STD | MSE | MAE | Median Bias |
|---|---|---|---|---|---|---|
| 5 | MILE | 0.0502 | 0.1204 | 0.0170 | 0.0906 | 0.0419 |
| | MILE(1) | −0.0074 | 0.0701 | 0.0050 | 0.0486 | −0.0068 |
| | MILE($\infty$) | −0.0090 | 0.0612 | 0.0038 | 0.0469 | −0.0068 |
| | AB(observed) | −0.2919 | 0.0473 | 0.0874 | 0.2919 | −0.2909 |
| | AB(expected) | −0.0849 | 0.0525 | 0.0100 | 0.0868 | −0.0822 |
| | AB(1)(expected) | −0.0249 | 0.0585 | 0.0040 | 0.0510 | −0.0245 |
| | IV | 0.0047 | 0.1446 | 0.0209 | 0.1131 | 0.0043 |
| | Lancaster | 0.0013 | 0.0648 | 0.0042 | 0.0524 | 0.0016 |
| | Arellano-Bond (GMM) | −0.0845 | 0.1402 | 0.0268 | 0.1295 | −0.0844 |
| | Arellano-Bond (Pooled) | −0.0760 | 0.2359 | 0.0614 | 0.1902 | −0.0699 |
| | MLE | −0.3322 | 0.0467 | 0.1125 | 0.3322 | −0.3317 |
| 10 | MILE | 0.0036 | 0.0376 | 0.0014 | 0.0303 | 0.0040 |
| | MILE(1) | 0.0008 | 0.0367 | 0.0013 | 0.0295 | 0.0016 |
| | MILE($\infty$) | 0.0008 | 0.0367 | 0.0013 | 0.0295 | 0.0016 |
| | AB(observed) | −0.1252 | 0.0356 | 0.0169 | 0.1252 | −0.1239 |
| | AB(expected) | −0.0315 | 0.0343 | 0.0022 | 0.0379 | −0.0303 |
| | AB(1)(expected) | −0.0071 | 0.0364 | 0.0014 | 0.0298 | −0.0069 |
| | IV | 0.0048 | 0.0773 | 0.0060 | 0.0611 | 0.0054 |
| | Lancaster | 0.0007 | 0.0376 | 0.0014 | 0.0301 | 0.0009 |
| | Arellano-Bond (GMM) | −0.1848 | 0.1177 | 0.0480 | 0.1876 | −0.1706 |
| | Arellano-Bond (Pooled) | −0.0493 | 0.1795 | 0.0346 | 0.1433 | −0.0385 |
| | MLE | −0.1617 | 0.0310 | 0.0271 | 0.1617 | −0.1606 |
| 20 | MILE | −0.0000 | 0.0223 | 0.0005 | 0.0178 | −0.0002 |
| | MILE(1) | −0.0001 | 0.0223 | 0.0005 | 0.0178 | −0.0003 |
| | MILE($\infty$) | −0.0001 | 0.0223 | 0.0005 | 0.0178 | −0.0003 |
| | AB(observed) | −0.0534 | 0.0238 | 0.0034 | 0.0535 | −0.0534 |
| | AB(expected) | −0.0095 | 0.0229 | 0.0006 | 0.0198 | −0.0093 |
| | AB(1)(expected) | −0.0014 | 0.0233 | 0.0005 | 0.0187 | −0.0006 |
| | IV | 0.0006 | 0.0447 | 0.0020 | 0.0358 | 0.0006 |
| | Lancaster | -0.0001 | 0.0234 | 0.0005 | 0.0187 | 0.0005 |
| | Arellano-Bond (GMM) | −0.3174 | 0.1061 | 0.1120 | 0.3174 | −0.3057 |
| | Arellano-Bond (Pooled) | −0.0436 | 0.1472 | 0.0235 | 0.1191 | −0.0305 |
| | MLE | −0.0787 | 0.0207 | 0.0066 | 0.0787 | −0.0793 |

TABLE 4. Empirical coverage probability of the LR-based confidence region for the logit model ($n = 100$).

| $T$ | MILE | MILE(1) | MILE($\infty$) | AB(observed) | AB(expected) | MLE | Lancaster |
|---|---|---|---|---|---|---|---|
| 5 | 0.9020 | 0.9010 | 0.9010 | 0.5100 | 0.7830 | 0.4420 | 0.8870 |
| 10 | 0.9360 | 0.9360 | 0.9360 | 0.8880 | 0.9100 | 0.7220 | 0.9360 |
| 20 | 0.9400 | 0.9400 | 0.9400 | 0.9280 | 0.9300 | 0.8450 | 0.9400 |

nominal coverage = 95%

TABLE 5. Empirical coverage probability of the LR-based confidence region for the probit model ($n = 100$).

| $T$ | MILE | MILE(1) | MILE($\infty$) | AB(observed) | AB(expected) | MLE | Lancaster |
|---|---|---|---|---|---|---|---|
| 5 | 0.8850 | 0.8740 | 0.8740 | 0.1230 | 0.6310 | 0.1330 | 0.8380 |
| 10 | 0.9490 | 0.9470 | 0.9470 | 0.7520 | 0.8600 | 0.4560 | 0.9400 |
| 20 | 0.9490 | 0.9490 | 0.9490 | 0.9150 | 0.9160 | 0.6930 | 0.9470 |

nominal coverage = 95%

TABLE 6. Empirical coverage probability of the LR-based confidence region for the AR(1) model ($n = 100$).

| $T$ | 5 | 10 | 20 |
|---|---|---|---|
| MILE | 0.8580 | 0.9460 | 0.9580 |
| MILE(1) | 0.9560 | 0.9510 | 0.9600 |
| MILE($\infty$) | 0.9670 | 0.9510 | 0.9600 |
| AB(observed) | 0.0000 | 0.0070 | 0.2430 |
| AB(expected) | 0.6190 | 0.8110 | 0.9010 |
| AB(1)(expected) | 0.8890 | 0.9080 | 0.9130 |
| Lancaster | 0.9440 | 0.9370 | 0.9330 |
| MLE | 0.0000 | 0.0000 | 0.0310 |

nominal coverage = 95%

## References


Amemiya, T. (1985): *Advanced econometrics*. Harvard University Press, Cambridge, MA, USA.

Andersen, E. B. (1970): "Asymptotic properties of conditional maximum-likelihood estimators," *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

Anderson, T. W., and C. Hsiao (1981): "Estimation of dynamic models with error components," *Journal of the American Statistical Association*, 76, 598–606.

Apostol, T. M. (1981): *Mathematical analysis*. Addison-Wesley, Reading, MA, USA, 2nd edn.

Arellano, M. (2003a): "Discrete choices with panel data," *Investigaciones Económicas*, XXVII, 423–458.

——— (2003b): *Panel data econometrics*. Oxford University Press, New York, NY, USA.

Arellano, M., and S. Bond (1991): "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, 58, 277–297.

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

Arellano, M., and J. Hahn (2006): "A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects," Manuscript, `www.cemfi.es/~arellano/arellano-hahn-paper2006.pdf`.

——— (2007): "Understanding bias in nonlinear panel models: Some recent developments," in *Advances in Economics and Econometrics: Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 3, pp. 381–409. Cambridge University Press, Cambridge, UK.

——— (2016): "A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects," *Global Economic Review*, 45, 251–274.

Baltagi, B. H. (1995): *Econometric analysis of panel data*. Wiley, Great Britain.

Barndorff-Nielsen, O. E. (1983): "On a formula for the distribution of the maximum likelihood estimator," *Biometrika*, 70, 343–365.

Bester, C. A., and C. Hansen (2009): "A penalty function approach to bias reduction in nonlinear panel models with fixed effects," *Journal of Business and Economic Statistics*, 27, 131–148.

Carro, J. M. (2007): "Estimating dynamic panel data discrete choice models with fixed effects," *Journal of Econometrics*, 140, 503–528.

Chamberlain, G. (1980): "Analysis of covariance with qualitative data," *Review of Economic Studies*, XLVII, 225–238.

——— (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5–46.

——— (1984): "Panel data," in *Handbook of Econometrics, vol. II*, ed. by Z. Griliches, and M. D. Intriligator, pp. 1247–1318. Elsevier, The Netherlands.

——— (2010): "Binary response models for panel data: Identification and information," *Econometrica*, 78, 159–168.

Cox, D. R., and N. Reid (1987): "Parameter orthogonality and approximate conditional inference," *Journal of the Royal Statistical Society, Series B*, 49, 1–39.

De Bin, R., N. Sartori, and T. A. Severini (2015): "Integrated likelihoods in models with stratum nuisance parameters," *Electronic Journal of Statistics*, 9, 1474–1491.

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

——— (2016): "Likelihood inference in an autoregression with fixed effects," *Econometric Theory*, 32, 1178–1215.

Dhaene, G., and Y. Sun (2020): "Second-order corrected likelihood for non-linear panel models with fixed effects," *Journal of Econometrics*, Forthcoming. https://doi.org/10.1016/j.jeconom.2020.04.001.

DiCiccio, T. J., M. A. Martin, S. E. Stern, and G. A. Young (1996): "Information bias and adjusted profile likelihoods," *Journal of the Royal Statistical Society, Series B*, 58, 189–203.

Durrett, R. (1991): *Probability*: *Theory and Examples*. Wadsworth & Brooks/Cole, Pacific Grove, CA, USA.

Ferguson, T. S. (1996): *A course in large sample theory*. Chapman & Hall, New York, NY, USA.

Fernández-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models," *Journal of Econometrics*, 150, 71–85.

Fernández-Val, I., and M. Weidner (2016): "Individual and time effects in nonlinear panel models with large $N, T$," *Journal of Econometrics*, 192, 291–312.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both $n$ and $T$ are large," *Econometrica*, 70, 1639–1657.

——— (2011): "Bias reduction for dynamic nonlinear panel models with fixed effects," *Econometric Theory*, 27, 1152–1191.

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

Hsiao, C. (1986): *Analysis of panel data*. Cambridge University Press, Cambridge, UK.

Kass, R. E., L. Tierney, and J. B. Kadane (1990): "The validity of posterior expansions based on Laplace's method," in *Bayesian and likelihood methods in statistics and econometrics*, ed. by S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner. Elsevier, The Netherlands.

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

——— (2002): "Orthogonal parameters and panel data," *Review of Economic Studies*, 69, 647–666.

Levin, B. (1990): "The saddlepoint correction in conditional logistic likelihood analysis," *Biometrika*, 77, 275–285.

Li, H., B. G. Lindsay, and R. P. Waterman (2003): "Efficiency of projected score methods in rectangular array asymptotics," *Journal of the Royal Statistical Society, Series B*, 65, 191–208.

Magnac, T. (2004): "Panel binary variables and sufficiency: Generalizing conditional logit," *Econometrica*, 72, 1859–1876.

McCullagh, P., and R. Tibshirani (1990): "A simple method for the adjustment of profile likelihoods," *Journal of the Royal Statistical Society, Series B*, 52, 325–344.

Mikailov, N. (2017): "Integrated likelihood estimation of panel data models with individual and time effects," Master's thesis in Economics and Finance, University of Luxembourg.

Newey, W. K. (1991): "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, 59, 1161–1167.

Neyman, J., and E. L. Scott (1948): "Consistent estimation from partially consistent observations," *Econometrica*, 16, 1–32.

Nickell, S. (1981): "Biases in dynamic models with fixed effects," *Econometrica*, 49, 1417–1426.

Pace, L., and A. Salvan (2006): "Adjustments of the profile likelihood from a new perspective," *Journal of Statistical Planning and Inference*, 136, 3554–3564.

Pakel, C. (2019): "Bias reduction in nonlinear and dynamic panels in the presence of cross-section dependence," *Journal of Econometrics*, 213, 459–492.

Schumann, M. (2020): "Second order bias reduction for nonlinear panel data models with fixed effects based on expected quantities," Manuscript, `https://sites.google.com/site/martinschumannecon/research`.

Schumann, M., T. A. Severini, and G. Tripathi (2020): "The role of information unbiasedness in panel data likelihoods," In progress.

Severini, T. A. (2000): *Likelihood methods in statistics*. Oxford University Press, London.

——— (2007): "Integrated likelihood functions for non-Bayesian inference," *Biometrika*, 94, 529–542.

Varrette, S., P. Bouvry, H. Cartiaux, and F. Georgatos (2014): "Management of an Academic HPC Cluster: The UL Experience," in *Proceedings of the* 2014 *International Conference on High Performance Computing and Simulation* (*HPCS* 2014). IEEE, Bologna, Italy.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, MA, USA, 2nd edn.

Woutersen, T. (2002): "Robustness against incidental parameters," Manuscript, University of Western Ontario, Department of Economics Research Reports, 2002-8. ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1391&context=economicsresrpt.

Zorich, V. A. (2004): *Mathematical Analysis I*. Springer-Verlag, Berlin.

Department of Quantitative Economics, School of Business and Economics, Maastricht University, 6211 LM Maastricht, The Netherlands.
*Email address*: m.schumann@maastrichtuniversity.nl

Department of Statistics, Northwestern University, Evanston, IL-60201, U.S.A.
*Email address*: severini@northwestern.edu

Department of Economics and Management, University of Luxembourg, L-1359, Luxembourg.
*Email address*: gautam.tripathi@uni.lu