Journal of Econometrics 235 (2023) 1215–1238

Contents lists available at ScienceDirect

Journal of Econometrics

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

The role of score and information bias in panel data likelihoods

Martin Schumann [a], Thomas A. Severini [b], Gautam Tripathi [c,*]

[a] Department of Quantitative Economics, Maastricht University, 6211 LM Maastricht, The Netherlands
[b] Department of Statistics, Northwestern University, Evanston, IL 60201, USA
[c] Department of Economics and Management, University of Luxembourg, L-1359, Luxembourg

ARTICLE INFO

Article history:
Received 20 February 2021
Received in revised form 3 August 2022
Accepted 9 August 2022
Available online 23 December 2022

JEL classification:
C23

Keywords:
Fixed effects
Hessian bias
Information bias
Panel data likelihood
Score bias

ABSTRACT

We show why reducing information bias can improve the performance of likelihood based estimators and confidence regions in small samples, and why it seems to matter more for inference than for estimation. The insights in this paper are helpful in explaining several simulation findings in the panel data literature. E.g., we can explain the well documented phenomenon that reducing the score bias alone often reduces the finite sample variance of estimators and improves the coverage of confidence regions in small samples, and why confidence regions based on conditional (on sufficient statistics) likelihoods can have excellent coverage even in very short panels. We can also explain the simulation results in Schumann, Severini, and Tripathi (2021), who find that, in short panels, estimators and confidence regions based on pseudolikelihoods that are simultaneously first-order score and information unbiased perform much better than those based on pseudolikelihoods that are only first-order score unbiased.

1. Introduction

Let θ denote a (column) vector of parameters of interest. In order to do likelihood based inference for θ in the presence of nuisance parameters, the typical first step is to eliminate the latter from the joint likelihood of θ and the nuisance parameters. This leads to a "pseudolikelihood" of θ,[1] which can be used to estimate θ and do inference based on the likelihood ratio statistic. Since nuisance parameters can be eliminated in at least three different ways, e.g., by profiling them out, by conditioning on sufficient statistics, or by integrating them out, how well a resulting pseudolikelihood of θ performs in practice depends on how good an approximation it is to a benchmark likelihood of θ. The benchmark likelihood, which we call the "target" likelihood, is a genuine oracle, i.e., infeasible, likelihood of θ.

The usual approach to determine whether a pseudolikelihood behaves like the target likelihood is to check if it satisfies the 1st Bartlett identity ("score unbiasedness") and the 2nd Bartlett identity ("information unbiasedness"). Since score unbiasedness determines the consistency of likelihood estimators (McCullagh and Tibshirani, 1990, Remark 2, p. 329), the literature has focused a lot of attention on constructing pseudolikelihoods possessing score bias of smaller order. These include, e.g., marginal and conditional likelihoods (Kalbfleisch and Sprott, 1970, 1973; Chamberlain, 1980), the modified profile likelihood (Barndorff-Nielsen, 1983), the Cox–Reid approximate conditional likelihood (Cox and Reid,

* Corresponding author.
E-mail addresses: m.schumann@maastrichtuniversity.nl (M. Schumann), severini@northwestern.edu (T.A. Severini), gautam.tripathi@uni.lu (G. Tripathi).

[1] A pseudolikelihood of θ is a nonnegative random function of θ that does not satisfy at least one of the Bartlett identities. In contrast, a "genuine" likelihood of θ is a nonnegative random function of θ that satisfies all of the Bartlett identities. Cf. Severini (2000, Section 3.5.2) for more on the Bartlett identities.

https://doi.org/10.1016/j.jeconom.2022.08.011
0304-4076/© 2022 Elsevier B.V. All rights reserved.

1987), score-bias-corrected profile likelihoods (McCullagh and Tibshirani, 1990; Arellano and Hahn, 2016), and integrated likelihoods (Kalbfleisch and Sprott, 1970; Berger et al., 1999; Lancaster, 2002; Arellano and Bonhomme, 2009). Cf. Severini (2000) and Arellano and Hahn (2007) for additional references.

In contrast, although it is reasonable to believe that likelihood-based methods should lead to even better inference when the information bias is also of smaller order – because when the score and information bias are both small, the pseudolikelihood should, at least in an intuitive sense, be closer to the target likelihood – the existing literature is not very precise about why reducing information bias can lead to improved inference in small samples. For instance, the earliest reference to the potential advantage of reducing information bias that we are aware of is McCullagh and Tibshirani (1990, Section 6). However, as they note explicitly on p. 326 of their paper, they have "…no strong argument to support this claim". In other works, this can be inferred from what is not said. E.g., the well-known DiCiccio et al. (1996) paper on reducing information bias never actually states why it should be reduced. Severini (1998, Section 4) considers information bias for pseudolikelihoods, again without explicitly saying why having no information bias is beneficial. Mykland (1999, Section 3) shows that there is some advantage to having bounded information bias, but he does not show any benefit to reducing the information bias further.

One reason why the aforementioned papers find it difficult to connect information bias reduction with inference may be that they restrict their attention to settings, e.g., pure cross-sectional likelihoods based on a large sample and a fixed number of parameters, where the source of randomness comes from a single index. The lack of variability along a second dimension means that in single index asymptotic expansions where the terms due to information bias appear, there tend to be a number of other components of the same order and, hence, simply reducing information bias does not reduce the order of the terms. In contrast, as we do in this paper, in a panel data setting where the number of individuals (n) and the length of the panel (T) are both allowed to grow, the presence of a second dimension can be exploited to characterize the effect of reducing information bias on statistical inference. We focus on short panels, relevant for microeconometric applications, where both $n, T \to \infty$ such that n grows as least as fast as T.

The idea that simultaneously reducing the score and information bias of panel data pseudolikelihoods can lead to better estimation and inference in small samples is supported by the simulation evidence in Schumann et al. (2021), henceforth referred to as SST. In the current paper, we provide a theoretical argument to show why this improved performance can be attributed to first-order score and information unbiasedness. Earlier works in the panel data literature that report improved inference for modified likelihood ratios do not investigate the source of the improvements. To the best of our knowledge, our paper is the first to make explicit the connection between the bias of the likelihood ratio and the score and information bias of the pseudolikelihood by showing that the improvement in the bias of the likelihood ratio resulting from reducing the score bias of the pseudolikelihood alone comes exclusively from changing an $O_p(n/T)$ term in its approximation to an $O_p(n/T^3)$ term, and that the improvement from reducing the information bias of the pseudolikelihood alone comes from reducing an $O_p(1/T)$ term to an $O_p(1/T^2)$ term (Lemma 6.1 and its discussion). Furthermore, as discussed after Lemma 6.2, we show how it – together with the concept of "hessian bias" introduced in Section 6.2 – can be used to explain why information bias seems to matter more for inference than estimation, and why reducing the pseudolikelihood score bias alone often reduces the variance of the resulting estimator and improves the coverage of likelihood ratio confidence regions in small samples. By connecting the target and conditional (on sufficient statistics) likelihoods (Section 5), we demonstrate that the results in Lemmas 6.1 and 6.2 also apply to conditional likelihoods.

The rest of the paper is organized as follows. Section 2 introduces the model, namely, a panel data likelihood with fixed effects, Section 3 defines the concepts of first-order score and information unbiasedness, and a generic panel data pseudolikelihood is described in Section 4. The target and conditional panel data likelihoods are discussed in Section 5. Section 6, which contains the main results of this paper, shows how the score, information, and hessian bias of a pseudolikelihood affects estimation and inference. Section 7 summarizes the relevance of our results for applied econometric practice, provides some motivation behind the proof of Lemma 6.1, and discusses its extension to nonlinear dynamic models. Section 8 concludes. The assumptions needed to derive the results, and their discussion, are in Appendix A. The {score, information, hessian}-bias of the pseudolikelihoods in the Neyman–Scott model are examined in Appendix B, and the proof of Lemma 6.1 is outlined in Appendix C. Lemma A.1, which suggests why Lemma 6.1 also holds for nonlinear dynamic AR(p) panel data models, is proved in Appendix D. Proofs of the remaining results are in Appendices E–H, available as supplementary material for this paper.

2. A panel data likelihood with fixed effects

Let Y_{it} denote outcomes and X_{it} a (column) vector of explanatory variables for $i = 1, \dots, n$ and $t = 1, \dots, T$, where $n, T \geq 2$. The random variables Y_{it} and X_{it} are observed, with i indexing the individual and t the time. The time-invariant fixed effect, denoted by α_{i0}, is an individual specific unobserved random variable whose distribution has known support supp(α_{i0}) $\subset \mathbb{R}$, but is otherwise unknown. Let $\mathcal{Y}_{iT} := (Y_{i1}, \dots, Y_{iT})$ denote the time-series of outcomes, and $\mathcal{X}_{iT} := (X_{i1}, \dots, X_{iT})$ the time-series of explanatory variables, corresponding to the ith individual for the length of the panel.

The joint distribution of $(\mathcal{X}_{iT}, \alpha_{i0})$ is unknown, which allows for arbitrary correlation between the explanatory variables and the fixed effect. 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 $\theta_0 \in \operatorname{int}(\Theta)$, where Θ is a known convex subset of $\mathbb{R}^{\dim(\theta_0)}$ with nonempty interior. It is assumed that, for each $(\theta, \alpha_i) \in \Theta \times \operatorname{supp}(\alpha_{i0})$, $f_{\mathcal{Y}_{iT} | \mathcal{X}_{iT}, \alpha_i; \theta}$ is a density with respect to an appropriate dominating measure (Lebesgue, counting, or a mixture of both).

Since α_{i0} is an unobserved random variable, we can talk about the likelihood of a potential realization. Specifically, if $\theta \in \Theta$ and $\alpha_i \in \operatorname{supp}(\alpha_{i0})$ denotes a possible value taken by α_{i0}, then we define the joint likelihood of (θ, α_i) for the ith individual to be $L_{iT}(\theta, \alpha_i) := L(\theta, \alpha_i; \mathcal{Y}_{iT}, \mathcal{X}_{iT}) := f_{\mathcal{Y}_{iT} | \mathcal{X}_{iT}, \alpha_i; \theta}(\mathcal{Y}_{iT})$. The average loglikelihood for the ith individual is denoted by $\ell_{iT}(\theta, \alpha_i) := T^{-1} \log L_{iT}(\theta, \alpha_i)$. We refer to θ as the parameter of interest, and call α_i an individual specific nuisance parameter.

The joint loglikelihood $(\theta, \alpha_i) \mapsto \ell_{iT}(\theta, \alpha_i)$ is assumed to be sufficiently well-behaved so that derivatives with respect to (θ, α_i), as many as needed, can be interchanged with integrals respect to $f_{\mathcal{Y}_{iT} | \mathcal{X}_{iT}, \alpha_i; \theta}$, and the mixed partial derivatives are equal. The score with respect to α_i is $\ell_{iT\alpha}(\theta, \alpha_i) := \partial_\alpha \ell_{iT}(\theta, \alpha_i)$, and $\ell_{iT\alpha\alpha}(\theta, \alpha_i) := \partial_\alpha^2 \ell_{iT}(\theta, \alpha_i)$.

3. Score and information bias of a panel data pseudolikelihood

Let $P_{iT}(\theta)$ denote a pseudolikelihood of θ for individual i after α_i has been eliminated from the joint likelihood $L_{iT}(\theta, \alpha_i)$ by using one of the methods described in the introduction.[2] The average pseudologlikelihood for the ith individual is $\mathfrak{p}_{iT}(\theta) := T^{-1} \log P_{iT}(\theta)$. The average pseudologlikelihood for the entire sample of individuals, denoted by $\mathfrak{p}_{\cdot T}(\theta) := n^{-1} \sum_{i=1}^n \mathfrak{p}_{iT}(\theta)$, can then be used to estimate θ and do inference. Specifically, the estimator of θ obtained by maximizing $\mathfrak{p}_{\cdot T}$ is given by $\hat{\theta}_{\mathfrak{p}} := \operatorname{argmax}_{\theta \in \Theta} \mathfrak{p}_{\cdot T}(\theta)$ (the dependence of $\hat{\theta}_{\mathfrak{p}}$ on n, T is suppressed to keep the notation simple), and the likelihood ratio of $\mathfrak{p}$, as a function of θ, is defined to be $\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta) := 2nT[\mathfrak{p}_{\cdot T}(\hat{\theta}_{\mathfrak{p}}) - \mathfrak{p}_{\cdot T}(\theta)]$.

Assume that $\hat{\theta}_{\mathfrak{p}}$ is consistent for θ_0 as $n, T \to \infty$, i.e., $\hat{\theta}_{\mathfrak{p}}$ converges in probability to θ_0 as $n, T \to \infty$. However, depending on the pseudolikelihood, e.g., the profile likelihood, the distribution of $\hat{\theta}_{\mathfrak{p}}$ may be asymptotically biased in the sense that the limiting distribution of $\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ may not be centered at zero as $n, T \to \infty$. If $\mathfrak{p}$ is such that the distribution of $\hat{\theta}_{\mathfrak{p}}$ is asymptotically unbiased, then the likelihood ratio of $\mathfrak{p}$ evaluated at θ_0, i.e., $\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0)$, is distributed as a $\chi^2_{\dim(\theta_0)}$ random variable as $n, T \to \infty$. This result can be used to test hypotheses and construct confidence regions for θ_0. E.g., it can be used to show that the lower-level random set $\{\theta \in \Theta : \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta) \le k_\tau\}$, where $\tau \in (0, 1)$ and k_τ denotes the $1 - \tau$ quantile of a $\chi^2_{\dim(\theta_0)}$ random variable, is a confidence region for θ_0 whose coverage probability approaches $1 - \tau$ as $n, T \to \infty$. However, if the distribution of $\hat{\theta}_{\mathfrak{p}}$ is asymptotically biased, then the limiting distribution of $\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0)$ is no longer $\chi^2_{\dim(\theta_0)}$ and, consequently, the likelihood ratio confidence region has poor empirical coverage even in large samples.

3.1. Score bias

A pseudolikelihood for individual i is said to be score unbiased, i.e., satisfy the 1st Bartlett identity, if $\mathbb{E}_0 \nabla_\theta \mathfrak{p}_{iT}(\theta_0) = 0$. Here, $\nabla_\theta := (\partial_\theta)'$ is the gradient, "$'$" the transpose operator, and – since $\mathfrak{p}_{iT}$ may depend on a preliminary estimator – $\mathbb{E}_0$ denotes expectation with respect to $\mathrm{pdf}_{\mathcal{Y}_{1T}, \ldots, \mathcal{Y}_{nT} | \mathcal{X}_{1T}, \alpha_{10}, \ldots, \mathcal{X}_{nT}, \alpha_{n0}} \overset{\text{Ass. A.1(i)}}{=} \prod_{i=1}^n f_{\mathcal{Y}_{iT} | \mathcal{X}_{iT}, \alpha_{i0}; \theta_0}$, the joint density of outcomes, given all the explanatory variables, for the entire sample of individuals.[3] Score unbiasedness generally does not hold for panel data pseudolikelihoods. Indeed, it is well known that eliminating the individual specific nuisance parameter from a panel data likelihood creates a bias, due to which it is typically the case that $\mathbb{E}_0 \nabla_\theta \mathfrak{p}_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-1})$ as $T \to \infty$. Since it is this bias that causes the limiting (as $n, T \to \infty$ at the same rate) distribution of $\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ to be incorrectly centered, a major focus of the literature has been to construct pseudolikelihoods for which the score bias is smaller by an order of magnitude, which ensures that the limiting distribution of $\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ is correctly centered when n grows at least as fast as T, but not too fast, e.g., $n/T^3 \to 0$. Consequently, if a pseudologlikelihood $\mathfrak{p}_{iT}$ satisfies the condition

$$\mathbb{E}_0 \nabla_\theta \mathfrak{p}_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-2}) \quad \text{as } T \to \infty, \tag{3.1}$$

then $\mathfrak{p}_{iT}$ is said to be "first-order score unbiased" (FOSU).[4] The left-hand side of (3.1), i.e., $\mathbb{E}_0 \nabla_\theta \mathfrak{p}_{iT}(\theta_0)$, is called the "score bias" of the pseudologlikelihood $\mathfrak{p}_{iT}$.

[2] Specific panel data pseudolikelihoods considered in this paper include the profile likelihood (Section 4), the conditional likelihood (Remark 5.1), and the integrated likelihoods of Lancaster (2002), Arellano and Bonhomme (2009), and SST, described briefly in Appendix G. These pseudolikelihoods have different properties because they differ in their construction. E.g., the integrated likelihoods of Arellano–Bonhomme and SST require a preliminary estimator in their construction; namely, Arellano–Bonhomme estimate the expectations appearing in their weight-functions by using time-series sample averages, whereas SST use the fixed effects maximum likelihood estimator (MLE) defined in Section 4 to construct their data based transformation of the fixed effects. In contrast, the profile likelihood, the conditional likelihood, and the integrated likelihood of Lancaster do not require preliminary estimators.

[3] If the pseudolikelihood does not depend on a preliminary estimator, then, by Assumption A.1(i), $\mathbb{E}_0 \nabla_\theta \mathfrak{p}_{iT}(\theta_0)$ reduces to an expectation with respect to the individual density $f_{\mathcal{Y}_{iT} | \mathcal{X}_{iT}, \alpha_{i0}; \theta_0}$. Since $\mathbb{E}_0$ is a conditional expectation, and var_0 the conditional variance using $\mathbb{E}_0$, where conditioning is on $(\mathcal{X}_{1T}, \alpha_{10}, \ldots, \mathcal{X}_{nT}, \alpha_{n0})$ or $(\mathcal{X}_{iT}, \alpha_{i0})$ depending on the context, equalities and inequalities involving them hold with probability one (w.p.1). This is the sense in which subsequent statements and assumptions regarding $\mathbb{E}_0$ and 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 var_0 is mentioned, we do not state them explicitly hereafter.

[4] If A_{iT} is an array, then the statement $A_{iT} = O_{\mathrm{p}}(1)$ is understood to hold element by element.

3.2. Information bias

A pseudolikelihood for individual i is said to be information unbiased, i.e., satisfy the 2nd Bartlett identity, if $T\mathbb{E}_0\nabla_\theta\mathfrak{p}_{iT}(\theta_0)\partial_\theta\mathfrak{p}_{iT}(\theta_0)+\mathbb{E}_0\nabla^2_{\theta\theta}\mathfrak{p}_{iT}(\theta_0)=0$,[5] where $\nabla^2_{ab}:=\partial_b\circ\nabla_a$. Eliminating the individual specific nuisance parameter creates a bias due to which information unbiasedness also does not hold for panel data pseudolikelihoods, and it is typically the case that $T\mathbb{E}_0\nabla_\theta\mathfrak{p}_{iT}(\theta_0)\partial_\theta\mathfrak{p}_{iT}(\theta_0)+\mathbb{E}_0\nabla^2_{\theta\theta}\mathfrak{p}_{iT}(\theta_0)=O_{\mathrm{p}}(T^{-1})$ as $T\to\infty$. Therefore, if a pseudologlikelihood $\mathfrak{p}_{iT}$ can be constructed such that its information bias is smaller by an order of magnitude, i.e., if it satisfies the condition

$$T\mathbb{E}_0\nabla_\theta\mathfrak{p}_{iT}(\theta_0)\partial_\theta\mathfrak{p}_{iT}(\theta_0)+\mathbb{E}_0\nabla^2_{\theta\theta}\mathfrak{p}_{iT}(\theta_0)=O_{\mathrm{p}}(T^{-2})\quad\text{as } T\to\infty, \tag{3.2}$$

then the pseudologlikelihood $\mathfrak{p}_{iT}$ is said to be "first-order information unbiased" (FOIU). The left-hand side of (3.2), i.e., $T\mathbb{E}_0\nabla_\theta\mathfrak{p}_{iT}(\theta_0)\partial_\theta\mathfrak{p}_{iT}(\theta_0)+\mathbb{E}_0\nabla^2_{\theta\theta}\mathfrak{p}_{iT}(\theta_0)$, is called the "information bias" of the pseudologlikelihood $\mathfrak{p}_{iT}$.

We investigate three mutually exclusive types of pseudolikelihoods: Those that are neither FOSU nor FOIU; those that are FOSU but not FOIU; those that are simultaneously FOSU and FOIU.[6] The most prominent example of a pseudolikelihood that is neither FOSU nor FOIU is the profile likelihood. Examples of pseudolikelihoods that are FOSU, but not FOIU, are the "information orthogonalizing transformation (IOT)" based integrated likelihood of Lancaster (2002), and the weighted integrated likelihood of Arellano and Bonhomme (2009). An example of a pseudolikelihood that is both FOSU and FOIU is the integrated likelihood of SST, which is based on the "zero-score-expectation (ZSE)" transformation of Severini (2007). Target and conditional likelihoods, being genuine likelihoods, are also FOSU and FOIU.

4. A generic panel data pseudolikelihood

The canonical example of a pseudolikelihood of θ is the profile loglikelihood $\ell^{\mathrm{p}}_{\cdot T}(\theta):=n^{-1}\sum_{i=1}^{n}\ell^{\mathrm{p}}_{iT}(\theta)$, where $\ell^{\mathrm{p}}_{iT}(\theta):=\ell_{iT}(\theta,\hat{\alpha}_{iT}(\theta))$ is the profile loglikelihood for individual i and $\hat{\alpha}_{iT}(\theta):=\operatorname{argmax}_{u\in\operatorname{supp}(\alpha_{i0})}\ell_{iT}(\theta,u)$ denotes the MLE of α_i for a given $\theta\in\Theta$. The fixed effects MLE of θ, which maximizes the profile likelihood for the entire sample of individuals, is denoted by $\tilde{\theta}:=\operatorname{argmax}_{\theta\in\Theta}\ell^{\mathrm{p}}_{\cdot T}(\theta)$. The profile likelihood is, in general, neither FOSU nor FOIU (DiCiccio, Martin, Stern, and Young, 1996, Section 3.1, p. 190).

The fact that the profile likelihood is not FOSU implies that although the fixed effects MLE is consistent as $n,T\to\infty$, its distribution is asymptotically biased in the sense that if n and T grow at the same rate, then $\sqrt{nT}(\tilde{\theta}-\theta_0)$ converges in distribution to a Gaussian random vector whose mean is not zero, cf., e.g., Li et al. (2003, Section 2) and Hahn and Newey (2004, Theorem 1). Consequently, the profile likelihood ratio $\mathrm{LR}^{\mathrm{p}}_{nT}(\theta_0):=2nT[\ell^{\mathrm{p}}_{\cdot T}(\tilde{\theta})-\ell^{\mathrm{p}}_{\cdot T}(\theta_0)]$ is no longer asymptotically distributed as a $\chi^2_{\dim(\theta_0)}$ random variable. It is, therefore, not surprising that in simulation studies, e.g., SST (Section 10), the profile likelihood ratio confidence regions are found to have poor coverage in small samples.

Research efforts to solve this problem, namely, to construct a pseudologlikelihood whose maximizer has a correctly centered limiting distribution with variance equal to that of the fixed effects MLE, have led to several competing candidates. The distinguishing feature of these competitors is that they can all be expressed as an additive correction, explicit or otherwise, to the profile loglikelihood. That is, each pseudologlikelihood $\mathfrak{p}_{iT}$ in the literature that claims to solve the problems besetting the profile loglikelihood can be written as

$$\mathfrak{p}_{iT}(\theta)=\begin{cases}\ell^{\mathrm{p}}_{iT}(\theta)+\mathfrak{C}_{iT}(\theta) & \text{if the correction is explicit}\quad\text{(a)}\\ \ell^{\mathrm{p}}_{iT}(\theta)+\mathfrak{C}_{iT}(\theta)+O_{\mathrm{p}}(T^{-2}) & \text{if the correction is implicit,}\quad\text{(b)}\end{cases} \tag{4.1}$$

where $\mathfrak{C}_{iT}(\theta)$ is the additive correction to $\ell^{\mathrm{p}}_{iT}(\theta)$. Though denoted by the same symbol, the additive corrections in 4a and 4b can be very different. An explicit additive correction means that one is imposed definitionally, i.e., $\mathfrak{p}_{iT}(\theta):=\ell^{\mathrm{p}}_{iT}(\theta)+\mathfrak{C}_{iT}(\theta)$; cf., e.g., McCullagh and Tibshirani (1990), Arellano (2003), Sartori (2003), Pace and Salvan (2006), Arellano and Hahn (2007, 2016), Schumann (2022), and the references therein. In contrast, an implicit (or internal) additive correction means that the definition of $\mathfrak{p}_{iT}$ can be used to deduce the existence of a function $\mathfrak{C}_{iT}$ such that $\mathfrak{p}_{iT}(\theta)-\ell^{\mathrm{p}}_{iT}(\theta)=\mathfrak{C}_{iT}(\theta)+O_{\mathrm{p}}(T^{-2})$. Examples of pseudologlikelihoods where an additive correction to the profile loglikelihood is implicit include the integrated likelihoods of Lancaster, Arellano–Bonhomme, and SST.

Whether corrected explicitly or implicitly, most of the pseudolikelihoods proposed in the literature are only FOSU. E.g., SST (Section 7.5) have shown that the integrated likelihoods of Lancaster and Arellano–Bonhomme are not FOIU. Although first-order score unbiasedness is enough to guarantee that the estimators proposed by Lancaster and Arellano–Bonhomme are well behaved as $n,T\to\infty$, simulation results in SST (Section 10) reveal that, in short panels: (i) The empirical coverage of the likelihood ratio confidence regions based on pseudolikelihoods that are simultaneously FOSU and FOIU is much higher than the empirical coverage of the likelihood ratio confidence regions based on pseudolikelihoods that are only FOSU; (ii) Estimators obtained by maximizing pseudolikelihoods that are simultaneously FOSU and FOIU have smaller finite sample variance than estimators obtained by maximizing pseudolikelihoods that are only FOSU. In Section 6, we provide a theoretical argument to explain why the source of this discrepancy can be attributed to first-order information unbiasedness.

[5] The factor T before the $\mathbb{E}_0\nabla_\theta\mathfrak{p}_{iT}(\theta_0)\partial_\theta\mathfrak{p}_{iT}(\theta_0)$ term is not a typo. It is due to the fact that $\mathfrak{p}_{iT}$ is defined to be the average (over T) pseudologlikelihood for individual i.

[6] Pseudolikelihoods that are FOIU but not FOSU can also exist. E.g., SST (Supplement, p. 123) have shown that the profile likelihood in the dynamic Neyman–Scott model (a linear AR(1) panel data model with Gaussian innovations) is FOIU but not FOSU. However, as noted in Remark 6.1(i), information bias reduction is pointless without score bias reduction. Therefore, we do not consider pseudolikelihoods that are FOIU but not FOSU.

5. The target likelihood

Following Pace and Salvan (2006, Section 3.2), the target loglikelihood of θ for the ith individual is $\ell_{iT}(\theta) := \ell_{iT}(\theta, \alpha^*_{iT}(\theta))$, where $\alpha^*_{iT}(\theta) := \operatorname{argmax}_{u \in \operatorname{supp}(\alpha_{i0})} \mathbb{E}_0 \ell_{iT}(\theta, u)$ is the "population level" MLE of α_i for a given θ. The average target loglikelihood for the entire sample of individuals, denoted by $\ell_{\cdot T}(\theta) := n^{-1} \sum_{i=1}^n \ell_{iT}(\theta)$, is an oracle, i.e., infeasible, loglikelihood because α^*_{iT} is unknown for each i, T. The target likelihood is a genuine likelihood because $L(\theta, \alpha^*_{iT}(\theta); \mathcal{Y}_{iT}, \mathcal{X}_{iT}) := f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT}, \alpha^*_{iT}(\theta);\theta}(\mathcal{Y}_{iT})$ integrates to one with respect to $\mathcal{Y}_{iT}$ (for each $\theta, \mathcal{X}_{iT}$) since $\alpha^*_{iT}(\theta)$ does not depend on $\mathcal{Y}_{iT}$. Consequently, the target loglikelihood $\ell_{iT}(\theta)$ satisfies all of the Bartlett identities. In particular, it satisfies (3.1) and (3.2) exactly, i.e., with the $O_p(T^{-2})$ terms on their right-hand side replaced by zero. Therefore, the target likelihood serves as a natural benchmark to evaluate the performance of other pseudolikelihoods. The oracle estimator of θ_0, which maximizes the target likelihood for the entire sample of individuals, is denoted by $\hat{\theta}^* := \operatorname{argmax}_{\theta \in \Theta} \ell_{\cdot T}(\theta)$, and the target likelihood ratio is $\mathrm{LR}^{\text{target}}_{nT}(\theta_0) := 2nT[\ell_{\cdot T}(\hat{\theta}^*) - \ell_{\cdot T}(\theta_0)]$.

Remark 5.1 (*Conditional Likelihoods*)**.** In panel data models where sufficient statistics for the fixed effects exist – e.g., exponential family models such as the Neyman–Scott model, panel logit, panel Poisson, and the panel negative binomial model – the fixed effects can be eliminated by conditioning on the sufficient statistics, which leads to a conditional likelihood. Maximizing the conditional likelihood yields the conditional maximum likelihood estimator (CMLE). Like the oracle estimator, cf. the discussion after Assumption A.3, the CMLE is consistent and asymptotically unbiased, i.e., its limiting distribution is correctly centered, as $n \to \infty$ and T is held fixed. As shown below, conditional likelihoods are also genuine likelihoods. Hence, suitably modifying their proofs, the results we obtain for the target likelihood can also be shown to hold for conditional likelihoods.

To see that conditional likelihoods are genuine likelihoods, let $S_{\mathcal{X}_{iT}}(\mathcal{Y}_{iT})$ be a sufficient statistic for α_i. Denote by μ_0 the dominating measure for $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_i;\theta}$. By the factorization theorem for sufficient statistics (Halmos and Savage, 1949, Corollary 1), there exist functions $g_{\mathcal{X}_{iT},\theta,\alpha_i}, h_{\mathcal{X}_{iT},\theta} \geq 0$ μ_0-a.s. with $\int_{\mathcal{Y}_{iT}} h_{\mathcal{X}_{iT},\theta}(y)\,\mu_0(dy) < \infty$ such that $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_i;\theta}(\mathcal{Y}_{iT}) = g_{\mathcal{X}_{iT},\theta,\alpha_i}(S_{\mathcal{X}_{iT}}(\mathcal{Y}_{iT}))h_{\mathcal{X}_{iT},\theta}(\mathcal{Y}_{iT})$ μ_0-a.s. Absorbing $\int_{\mathcal{Y}_{iT}} h_{\mathcal{X}_{iT},\theta}(y)\,\mu_0(dy)$ in $g_{\mathcal{X}_{iT},\theta,\alpha_i}$, it can be assumed without loss of generality that $h_{\mathcal{X}_{iT},\theta}$ is a density with respect to μ_0. Therefore, as the conditional likelihood $L^{\text{cond}}(\theta; \mathcal{Y}_{iT}, \mathcal{X}_{iT}) := \mathbb{1}(f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_i;\theta}(\mathcal{Y}_{iT}) > 0)\frac{f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_i;\theta}(\mathcal{Y}_{iT})}{g_{\mathcal{X}_{iT},\theta,\alpha_i}(S_{\mathcal{X}_{iT}}(\mathcal{Y}_{iT}))} = h_{\mathcal{X}_{iT},\theta}(\mathcal{Y}_{iT})$ integrates to 1 with respect to $\mu_0(d\mathcal{Y}_{iT})$, it is a genuine likelihood. □

6. How do the score and information bias of a panel data pseudolikelihood affect estimation and inference of the parameter of interest?

6.1. The effect of first-order score and information unbiasedness of the pseudolikelihood on the bias of the likelihood ratio

The conditional bias (given the covariates $\mathcal{X}_{1T}, \alpha_{10}, \ldots, \mathcal{X}_{nT}, \alpha_{n0}$) of the likelihood ratio, denoted by $\mathbb{E}_0 \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) - \dim(\theta_0)$, is related to the conditional coverage probability of the likelihood ratio confidence region. Specifically, the coverage probability approaches its nominal value as the expectation of the likelihood ratio gets closer to its nominal value.[7] Therefore, we now demonstrate how the score and information bias of $\mathfrak{p}_{iT}$ asymptotically affect $\mathbb{E}_0 \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) - \dim(\theta_0)$. To motivate this, we begin by looking at a famous example due to Neyman and Scott (1948).

Example 6.1 (*Heterogeneous means model of Neyman–Scott*)**.** Consider a Gaussian panel data model where individual specific heterogeneity arises only in the mean, i.e., let $Y_{it} = \alpha_{i0} + U_{it}$, where $U_{i1}, \ldots, U_{iT} \mid \alpha_{i0} \overset{\mathrm{d}}{=} \mathrm{NIID}(0, \sigma_0^2)$ with $\operatorname{supp}(\alpha_{i0}) = \mathbb{R}$ and $\sigma_0^2 \in (0, \infty) =: \Theta$. The Neyman–Scott model has been used to explain variation in scores $(Y_{i1}, \ldots, Y_{iT})$ obtained by individual i on repeated IQ tests using unobserved individual ability (α_{i0}) as the only explanatory variable. The parameter of interest is the variance of the IQ scores, i.e., $\theta := \sigma^2$. In this example, the fixed effects MLE of θ is not consistent for θ_0 as $n \to \infty$ and T is held fixed. Although the consistency issue can be fixed by letting T grow with n, it can be shown (SST, Supplement, Example B.2) that the asymptotic distribution of the fixed effects MLE of σ^2 is not correctly centered when n grows at least as fast as T. This is the classic incidental parameters problem of maximum likelihood, first noted by Neyman and Scott (1948).

[7] To get some intuition behind this phenomenon, suppose that $n = 1$ and there are no nuisance parameters and covariates. Let $\mathrm{LR}_T := \mathrm{LR}_T(\theta_0)$ be the likelihood ratio evaluated at θ_0; its expectation $\mu_T := \mathbb{E}\mathrm{LR}_T$ now only depends on T. It is known from the theory of Bartlett correction of the likelihood ratio (Barndorff-Nielsen and Hall, 1988) that $\mu_T = \dim(\theta_0) + O(T^{-1})$ and $\Pr(\dim(\theta_0)\mathrm{LR}_T/\mu_T \leq u) = F(u) + O(T^{-2})$, $u \in \mathbb{R}$, as $T \to \infty$, where F is the cumulative distribution function (c.d.f.) of a $\chi^2_{\dim(\theta_0)}$ random variable. Hence,

$$\Pr(\mathrm{LR}_T \leq u) = \Pr(\frac{\dim(\theta_0)\mathrm{LR}_T}{\mu_T} \leq \frac{\dim(\theta_0)u}{\mu_T}) = F(\frac{\dim(\theta_0)u}{\mu_T}) + O(\frac{1}{T^2}) = F(u) + O(\mu_T - \dim(\theta_0)) + O(\frac{1}{T^2}).$$

Therefore, $\mu_T - \dim(\theta_0)$, the bias of the likelihood ratio, helps determine the accuracy of the χ^2-approximation to the c.d.f. of the likelihood ratio. This suggests that the same holds for panel data pseudolikelihoods as well, although it is beyond the scope of our paper to establish this result rigorously.

The Neyman–Scott model provides a nice illustration of the effects of reducing the score and information bias of a pseudolikelihood on the bias of the likelihood ratio. To demonstrate this, we consider the likelihood ratios of three pseudolikelihoods and, for the sake of comparison, that of the target likelihood. The three pseudologlikelihoods we consider are the profile loglikelihood (ℓ^{p}_{iT}), the integrated loglikelihood of SST ($\bar{\ell}^{\mathrm{zse}}_{iT}$), and the explicitly corrected (or adjusted) pseudologlikelihood $\mathfrak{p}^{\mathrm{adj}}_{iT}(\theta) := \ell^{\mathrm{p}}_{iT}(\theta) + \mathfrak{c}_{iT}(\theta)$ with $\mathfrak{c}_{iT}$ defined in (B.16).[8]

It is shown in Appendix B that, in the Neyman–Scott model, $\bar{\ell}^{\mathrm{zse}}_{iT}$ is score as well as information unbiased, $\mathfrak{p}^{\mathrm{adj}}_{iT}$ is score unbiased but not FOIU, and that ℓ^{p}_{iT} is neither FOSU nor FOIU. Motivated by the arguments in Schumann (2022), the correction term in $\mathfrak{p}^{\mathrm{adj}}_{iT}$ is chosen such that the global maximizer of $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ coincides with the global maximizer of $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ (cf. the proof of (6.1)b). This ensures that any differences between the likelihood ratios of $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ and $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ can be attributed solely to the differences between the pseudolikelihoods themselves and not to the estimators used to construct the likelihood ratios.[9] As shown in Appendix E, the bias of the likelihood ratio in the Neyman–Scott model can be decomposed as follows (the limits in (6.1)a–(6.1)d are taken as $n \to \infty$, T-fixed):[10]

$$\mathbb{E}_0 \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) - 1 = \begin{cases} O(\frac{1}{nT}) + \underbrace{0}_{\text{due to score bias}} + \underbrace{0}_{\text{due to info bias}} & \text{if } \mathfrak{p} = \bar{\ell}^{\mathrm{zse}}_{\cdot T} \quad \text{(a)} \\ O(\frac{1}{nT}) + \underbrace{0}_{\text{due to score bias}} + \underbrace{O(\frac{1}{T})}_{\text{due to info bias}} & \text{if } \mathfrak{p} = \mathfrak{p}^{\mathrm{adj}}_{\cdot T} \quad \text{(b)} \\ \underbrace{O(\frac{n}{T})}_{\text{due to score bias}} + \underbrace{O(\frac{1}{T})}_{\text{due to info bias}} & \text{if } \mathfrak{p} = \ell^{\mathrm{p}}_{\cdot T} \quad \text{(c)} \\ O(\frac{1}{nT}) + \underbrace{0}_{\text{due to score bias}} + \underbrace{0}_{\text{due to info bias}} & \text{if } \mathfrak{p} = \text{target.} \quad \text{(d)} \end{cases} \tag{6.1}$$

Eqs. (6.1)a and (6.1)d show that, asymptotically, the likelihood ratio of $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ has the same bias as the likelihood ratio of the target likelihood, and the bias converges to zero as $n \to \infty$ and T is held fixed. In contrast, (6.1)b and (6.1)c reveal that the bias of the likelihood ratio of $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ converges to zero when T is allowed to grow with n, and the profile likelihood ratio bias vanishes only if T grows faster than n, which is not the right setting for modeling short panels. For $\mathfrak{p} \in \{\bar{\ell}^{\mathrm{zse}}_{\cdot T}, \mathfrak{p}^{\mathrm{adj}}_{\cdot T}\}$, the likelihood ratio bias is smallest when $\mathfrak{p} = \bar{\ell}^{\mathrm{zse}}_{\cdot T}$. As $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ and $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ are both score unbiased with the same global maximizer, this can be attributed to the fact that $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ is information unbiased whereas $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ is not. Indeed, their proofs reveal that the $O(1/T)$ term in (6.1)b, which is absent in (6.1)a, arises because $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ is not FOIU. The $O(n/T)$ term in (6.1)c, which is due to the fact that the profile likelihood is not FOSU, reveals that the bias of the profile likelihood ratio can be large when T is small, even if n is large. In (6.1)d, there are no terms due to score and information bias because the target likelihood is a genuine likelihood and, hence, satisfies all of the Bartlett identities. □

The results for the Neyman–Scott model can be generalized considerably to include nonlinear panel data models. The following proposition shows that, asymptotically, the bias of the likelihood ratio of a pseudologlikelihood $\mathfrak{p}_{iT}$ can be decomposed as the sum of two terms: One caused by the score bias of $\mathfrak{p}_{iT}$, and the other caused by the information bias of $\mathfrak{p}_{iT}$.

Lemma 6.1. *Let Assumptions A.1–A.7 hold. Then, as $n, T \to \infty$,*

$$\mathbb{E}_0 \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) - \dim(\theta_0) = \begin{cases} \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\textit{due to score bias}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\textit{due to info bias}} & \textit{if } \mathfrak{p} \textit{ is FOSU and FOIU} \quad \textit{(a)} \\ \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\textit{due to score bias}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\textit{due to info bias}} & \textit{if } \mathfrak{p} \textit{ is FOSU but not FOIU} \quad \textit{(b)} \\ \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{\textit{due to score bias}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\textit{due to info bias}} & \textit{if } \mathfrak{p} \textit{ is neither FOSU nor FOIU} \quad \textit{(c)} \\ O_{\mathrm{p}}(\frac{1}{nT}) & \textit{if } \mathfrak{p} \textit{ is the target loglikelihood.} \quad \textit{(d)} \end{cases} \tag{6.2}$$

[8] The reason we do not consider the integrated likelihoods of Lancaster and Arellano–Bonhomme is because in the Neyman–Scott model the integrated likelihoods of Lancaster and Arellano–Bonhomme – which are identical because the parameters are orthogonal – coincide with the integrated likelihood of SST constructed with $\pi_i := 1$ (Appendix G). This is not the case in general. It happens here only because the ZSE transformation in the Neyman–Scott model is the identity map (SST, Example 9.1).

[9] In the Neyman–Scott model, the maximum integrated likelihood estimator (MILE) proposed by SST and the oracle estimator are consistent for θ_0 as $n \to \infty$ and T is held fixed, whereas the fixed effects MLE, which maximizes the profile likelihood, is consistent for θ_0 only when both $n, T \to \infty$.

[10] In the Neyman–Scott model, the only explanatory variables are the fixed effects. Therefore, $\mathbb{E}_0 \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0)$ denotes expectation with respect to $\prod_{i=1}^{n} f_{\mathfrak{y}_{iT}|\mathfrak{x}_{iT},\alpha_{i0};\theta_0} = \prod_{i=1}^{n} \mathrm{pdf}_{U_{i1},\dots,U_{iT}|\alpha_{i0}} = \prod_{i=1}^{n}\prod_{t=1}^{T} \mathrm{pdf}_{\mathrm{N}(0,\sigma_0^2)}$.

The limits in Lemma 6.1 are taken as $n, T \to \infty$ because estimators in nonlinear panel data models are generally only consistent as $n, T \to \infty$. The results in (6.2)a–(6.2)d are qualitatively very similar to those in (6.1)a–(6.1)d, even though we are no longer in the Neyman–Scott setting. The terms due to the score bias depend on both n and T, whereas the terms due to the information bias depend only on T. The $O_p(T^{-2})$ term appears on the right-hand side only when the pseudolikelihood is FOIU.[11] If the pseudolikelihood is not FOIU, then the $O_p(T^{-2})$ term becomes an $O_p(T^{-1})$ term. The $O_p(n/T^3)$ term arises if the pseudolikelihood is FOSU.[12] If the pseudolikelihood is not FOSU, then the $O_p(n/T^3)$ term becomes $O_p(n/T)$. The $O_p(n/T)$ term in (6.2)c does not vanish because T does not grow faster than n (Assumption A.4(i)). This explains why the profile likelihood ratio behaves very poorly in nonlinear panel data models when T is small, even if n happens to be large.

Lemma 6.1 can be used to justify the improvements that researchers often discover in their simulation experiments when they compare the coverage probability of a likelihood ratio confidence region based on a FOSU pseudolikelihood with one based on the profile likelihood. Indeed, the improvement in going from (6.2)c to (6.2)b is caused by the $O_p(n/T)$ term in (6.2)c becoming the $O_p(n/T^3)$ term in (6.2)b. Similarly, comparing the $O_p(1/T^2)$ term in (6.2)a with the $O_p(1/T)$ term in (6.2)b suggests that the coverage probability of a confidence region obtained by inverting the likelihood ratio of a pseudolikelihood that is both FOSU and FOIU is likely to be higher than the coverage probability of a confidence region based on a pseudolikelihood that is only FOSU. Therefore, it makes sense to base inference on a pseudolikelihood that is both FOSU and FOIU.

Since the terms due to the score bias depend on both n and T, whereas the terms due to the information bias depend only on T, increasing n alone affects the score bias terms but not the information bias terms. If n and T grow at the same rate, an assumption maintained in several papers, then $O_p(n/T^3) = O_p(1/T^2)$ and the improvement to the coverage probability in (6.2)a from first-order information unbiasedness should be comparable to the improvement from first-order score unbiasedness. But if n grows too fast, say, $n = O(T^2)$ (so that the $O_p(n/T^3)$ term still vanishes asymptotically), then the $O_p(n/T^3)$ term in (6.2)a becomes $O_p(1/T)$, and an improvement to the coverage probability from first-order information unbiasedness is dominated by the improvement from first-order score unbiasedness. However, the simulation evidence in SST suggests that improvements to coverage probability in finite samples exist even when n is much larger than T. Therefore, using a pseudolikelihood that is simultaneously FOSU and FOIU is always beneficial.

Remark 6.1. (i) Lemma 6.1 makes it clear that there is no gain from first-order information unbiasedness without first-order score unbiasedness. Indeed, if $\mathfrak{p}$ were FOIU but not FOSU, then the $O_p(1/T)$ on the right-hand side of (6.2)c would be replaced by an $O_p(1/T^2)$ term, i.e.,

$$\mathbb{E}_0 \mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) - \dim(\theta_0)\big|_{\mathfrak{p}\text{ is FOIU but not FOSU}} = \underbrace{O_p\Big(\frac{n}{T}\Big)}_{\text{due to score bias}} + \underbrace{O_p\Big(\frac{1}{T^2}\Big)}_{\text{due to info bias}} = O_p\Big(\frac{n}{T}\Big).$$

(ii) Since conditional likelihoods are genuine likelihoods, the proof of (6.2)d can be adapted to show that it also holds for conditional likelihoods. E.g., this is confirmed in the Neyman–Scott model, where the integrated likelihood of SST coincides with the conditional likelihood. This explains why the likelihood ratio confidence regions based on conditional likelihoods can have excellent coverage probabilities even in very short panels. □

Example 6.2 (*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 ith individual is $L_{iT}(\theta, \alpha_i) = (\prod_{t=1}^T Y_{it}!)^{-1} e^{-\sum_{t=1}^T e^{X_{it}'\theta + \alpha_i}} e^{\sum_{t=1}^T Y_{it}(X_{it}'\theta + \alpha_i)}$. Lancaster (2002, p. 650) showed that $L_{iT}(\theta, \hat{\alpha}_{iT}(\theta)) \propto \mathrm{pmf}_{\mathcal{Y}_{iT}|\mathcal{X}_{iT}, \sum_{t=1}^T Y_{it}; \theta}(\mathcal{Y}_{iT})$, i.e., the profile likelihood for individual i coincides with the conditional density of $\mathcal{Y}_{iT}$ given $\mathcal{X}_{iT}$ and the statistic $\sum_{t=1}^T Y_{it}$, which is sufficient for α_i. SST (Example 9.2) have shown that the same also holds for their ZSE transformed integrated likelihood. Therefore, in the panel Poisson model, the profile likelihood, the ZSE transformed integrated likelihood, and the conditional likelihood, all coincide and, following Remark 6.1(ii), satisfy (6.2)d. □

6.2. *The effect of first-order score and information unbiasedness on the bias and variance of $\hat{\theta}_{\mathfrak{p}}$*

The following result shows how $\mathfrak{p}$ being FOSU or FOIU affects the conditional (on the covariates $\mathcal{X}_{1T}, \alpha_{10}, \ldots, \mathcal{X}_{nT}, \alpha_{1n}$) bias and variance of $\hat{\theta}_{\mathfrak{p}}$.

[11] There is no $O_p(T^{-2})$ term in (6.1)a because $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ is information unbiased in Example 6.1.

[12] There is no $O_p(n/T^3)$ term in (6.1)a and (6.1)b because, in Example 6.1, $\bar{\ell}^{\mathrm{zse}}_{\cdot T}$ and $\mathfrak{p}^{\mathrm{adj}}_{\cdot T}$ are score unbiased. Unlike (6.1)a and (6.1)b, no $O_p(1/nT)$ term appears in (6.2)a and (6.2)b because it is dominated by the $O_p(n/T^3)$ term, cf. (C.6) in the proof of (6.2)a.

Lemma 6.2. *Let Assumptions A.1–A.7 hold, and $n, T \to \infty$. Then,*

$$\mathbb{E}_0\hat{\theta}_{\mathfrak{p}} - \theta_0 = \begin{cases} \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{due\ to\ score\ bias} & \text{if } \mathfrak{p} \text{ is FOSU} \quad (a) \\ \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{due\ to\ score\ bias} & \text{if } \mathfrak{p} \text{ is not FOSU} \quad (b) \\ O_{\mathrm{p}}(\frac{1}{nT}) & \text{if } \mathfrak{p} \text{ is the target loglikelihood.} \quad (c) \end{cases} \tag{6.3}$$

Furthermore, if the expected hessian matrix $H_{\mathfrak{p}.T} := n^{-1}\sum_{i=1}^{n}\mathbb{E}_0\nabla^2_{\theta\theta}\mathfrak{p}_{iT}(\theta_0) =: n^{-1}\sum_{i=1}^{n}H_{\mathfrak{p}_{iT}}$, *and* var_0 *is the variance using* $\mathbb{E}_0$, *then*

$$\mathrm{var}_0\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0) + H^{-1}_{\mathfrak{p}.T} = \begin{cases} \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{due\ to\ score\ bias} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{due\ to\ info\ bias} & \text{if } \mathfrak{p} \text{ is FOSU and FOIU} \quad (a) \\ \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{due\ to\ score\ bias} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{due\ to\ info\ bias} & \text{if } \mathfrak{p} \text{ is FOSU but not FOIU} \quad (b) \\ \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{due\ to\ score\ bias} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{due\ to\ info\ bias} & \text{if } \mathfrak{p} \text{ is neither FOSU nor FOIU} \quad (c) \\ O_{\mathrm{p}}(\frac{1}{nT}) & \text{if } \mathfrak{p} \text{ is the target loglikelihood.} \quad (d) \end{cases} \tag{6.4}$$

Lemma 6.2 reveals some interesting findings. For instance, (6.3)a and (6.3)b show that, asymptotically, the bias of $\hat{\theta}_{\mathfrak{p}}$ depends only on T, and that its magnitude is determined solely by whether $\mathfrak{p}$ is FOSU or not. In particular, $\mathfrak{p}$ being FOIU does not reduce the bias of $\hat{\theta}_{\mathfrak{p}}$. Eq. (6.3)c shows that the bias of the oracle estimator goes to zero as $n \to \infty$ and T is held fixed, which is not surprising because the oracle estimator can be consistent for θ_0 as $n \to \infty$ and T is held fixed; cf. the discussion after Assumption A.3.

To assess the effect of first-order score and information unbiasedness of $\mathfrak{p}$ on the variance of estimators, we rewrite the inverse hessian matrix $H^{-1}_{\mathfrak{p}.T}$ in (6.4)a–(6.4)d in order to facilitate their comparison for different pseudologlikelihoods. Indeed, since the asymptotic variance of $\hat{\theta}_{\mathfrak{p}}$ is equal to the asymptotic variance of the oracle estimator, it is reasonable to express $-H_{\mathfrak{p}_{iT}}$ as a deviation from $F_{iT} := -H_{\mathfrak{p}_{iT}}\big|_{\mathfrak{p}=\text{target}} = -\mathbb{E}_0\nabla^2_{\theta\theta}\ell_{iT}(\theta_0, \alpha^*_{iT}(\theta_0))$, the Fisher information of the target likelihood for individual i.[13] To derive an expression for $H_{\mathfrak{p}_{iT}}$ in terms of its deviation from F_{iT}, recall from Section 4 that any panel data pseudologlikelihood $\mathfrak{p}_{iT}$ that claims to better the profile loglikelihood can be written as

$$\mathfrak{p}_{iT}(\theta) = \ell^{\mathrm{p}}_{iT}(\theta) + \mathcal{C}_{iT}(\theta) + O_{\mathrm{p}}(T^{-2}) \quad \text{as } T \to \infty, \tag{6.5}$$

where $\mathcal{C}_{iT}$ is a correction term designed to improve specific properties of the profile loglikelihood (if the correction is explicit, then the $O_{\mathrm{p}}(T^{-2})$ term in (6.5) is identically zero). E.g., several corrections exist that remove $\mathcal{B}_{iT}$, the term in the profile loglikelihood responsible for its first-order score bias (cf. (6.11)). Therefore, motivated by the fact that $\nabla^{\{0,1,2\}}_{\theta}\mathcal{B}_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-1})$ (cf. the proof of (6.6)), we assume henceforth that $\nabla^{\{0,1,2\}}_{\theta}\mathcal{C}_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-1})$.

It is shown in Appendix F.3 that if $\mathbb{E}_0\nabla^2_{\theta}\mathcal{C}_{iT}(\theta_0) = O_{\mathrm{p}}(T^{-1})$, then, under time-independence and mild conditions on some derivatives of the target loglikelihood at θ_0,

$$-H_{\mathfrak{p}_{iT}} = F_{iT} + \underbrace{O_{\mathrm{p}}(T^{-1})}_{\text{hessian bias}} \quad \text{as } T \to \infty, \tag{6.6}$$

where "hessian bias" refers to the approximation error when $-H_{\mathfrak{p}_{iT}}$ is approximated by F_{iT}, i.e., the hessian bias of a pseudolikelihood $\mathfrak{p}_{iT}$ is defined to be $-H_{\mathfrak{p}_{iT}} - F_{iT}$. Therefore, analogous to the definitions of first-order score and information unbiasedness in (3.1) and (3.2), we call a pseudolikelihood $\mathfrak{p}_{iT}$ to be "first-order Hessian unbiased" (FOHU) whenever

$$-H_{\mathfrak{p}_{iT}} = F_{iT} + O_{\mathrm{p}}(T^{-2}) \quad \text{as } T \to \infty. \tag{6.7}$$

Since the target likelihood satisfies (6.7) exactly, i.e., with the $O_{\mathrm{p}}(T^{-2})$ terms on its right-hand side replaced by zero, the target likelihood is, by definition, hessian unbiased.

Letting $F_{.T} := n^{-1}\sum_{i=1}^{n}F_{iT}$ denote average Fisher information of the target likelihood for the entire sample of individuals, we have $-H^{-1}_{\mathfrak{p}.T} \overset{(6.6)}{=} F^{-1}_{.T} + O_{\mathrm{p}}(T^{-1})$ under Assumption A.2. Hence, (6.4)a, (6.4)b, and (6.4)c imply that the

[13] The Fisher information, which is usually defined as the variance of the score function, is equal to the negative of the expected Hessian because the target likelihood satisfies the first two Bartlett identities.

difference between $\operatorname{var}_0 \sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ and $F_{\cdot T}^{-1}$ is given by

$$\operatorname{var}_0 \sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0) - F_{\cdot T}^{-1} = \begin{cases} \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{due to hessian bias}} + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\text{due to score bias}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{due to info bias}} & \textit{if } \mathfrak{p} \textit{ is } \text{FOSU } \textit{and } \text{FOIU} \quad \text{(a)} \\ \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{due to hessian bias}} + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\text{due to score bias}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{due to info bias}} & \textit{if } \mathfrak{p} \textit{ is } \text{FOSU } \textit{but not } \text{FOIU} \quad \text{(b)} \\ \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{due to hessian bias}} + \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{\text{due to score bias}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{due to info bias}} & \textit{if } \mathfrak{p} \textit{ is neither } \text{FOSU } \textit{nor } \text{FOIU.} \quad \text{(c)} \end{cases} \tag{6.8}$$

A well-documented phenomenon in the panel data literature — cf., e.g., the simulation results in Hahn and Newey (2004), Fernández-Val (2009), and SST — is that in simulations of fixed effects logit and probit models, estimators that correct the bias of the pseudolikelihood score not only perform better in terms of bias, but also have a lower finite sample variance than the fixed effects MLE.[14] One explanation for this finding is provided by Hahn and Newey (2004, p. 1307), who state that if the score bias can be attributed to the (normalized) scale parameter, then reducing the bias may lead to a decrease in the variance. An alternative explanation is given by (6.8)c and (6.8)b: Reducing the score bias causes the $O_{\mathrm{p}}(n/T)$ term in (6.8)c to be replaced by the $O_{\mathrm{p}}(n/T^3)$ term in (6.8)b, thereby reducing the variance.

In the decomposition for $\operatorname{var}_0 \sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0) - F_{\cdot T}^{-1}$, the term due to the hessian bias of $\mathfrak{p}$ dominates the term due to the information bias of $\mathfrak{p}$ by an order of magnitude when $\mathfrak{p}$ is FOIU. In contrast, the hessian bias of a pseudologlikelihood does not enter (6.2)a–(6.2)d at all. This explains why first-order information unbiasedness of a pseudologlikelihood has a greater impact on inference than on estimation. Unlike for the estimation bias, (6.8)a, (6.8)b, and (6.8)c suggest that $\mathfrak{p}$ being FOIU can reduce the deviation of the variance of $\hat{\theta}_{\mathfrak{p}}$ from the variance of the oracle estimator: The deviation is the largest if $\mathfrak{p}$ is neither FOSU nor FOIU, smaller if $\mathfrak{p}$ is FOSU but not FOIU, and the smallest if $\mathfrak{p}$ is both FOSU and FOIU. Since this ranking is based on comparing the rates inside the O_{p} terms in (6.8)a, (6.8)b, and (6.8)c, it may not hold if the constants in the O_{p} terms are not comparable across the pseudolikelihoods; cf., e.g., the discussion after SST (Eqn. 10.1).[15]

In some special cases, the variances of $\hat{\theta}_{\mathfrak{p}}$ and the oracle estimator can be obtained in closed form for each n, T without any O_{p} terms. E.g., it is shown in Example 6.3 that, in the Neyman–Scott model, the $O_{\mathrm{p}}(n/T)$ term in (6.8)c, which is typically the dominating term since n grows faster than T, does not appear in (6.10)c, and the $O_{\mathrm{p}}(n/T^3)$ terms in (6.8)a and (6.8)b are not present in (6.10)a and (6.10)b. Consequently, in Example 6.3, it is the fixed effects MLE that has the smallest finite sample variance of all the estimators considered there including the oracle estimator, and, hence, the smallest deviation from the oracle variance.

Example 6.3 (*Example* 6.1 *Contd.*)**.** For the pseudolikelihoods considered in Example 6.1, the baseline variance $F_{\cdot T}^{-1}$ and $\operatorname{var}_0 \sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ can be obtained in closed form. Indeed, as shown in Appendix E.3, in the Neyman–Scott model, we have, for each $n, T \geq 2$,

$$F_{\cdot T}^{-1} = 2\theta_0^2, \tag{6.9}$$

and

$$\operatorname{var}_0 \sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0) - F_{\cdot T}^{-1} = \begin{cases} \underbrace{\frac{2\theta_0^2}{T-1}}_{\text{due to hessian bias}} & \text{if } \mathfrak{p} = \bar{\ell}^{\text{zse}}_{\cdot T} \quad \text{(a)} \\ \underbrace{\frac{2\theta_0^2}{T-1}}_{\text{due to info bias}} & \text{if } \mathfrak{p} = \mathfrak{p}^{\text{adj}}_{\cdot T} \quad \text{(b)} \\ \underbrace{-\frac{2\theta_0^2}{T}}_{\text{due to score, info, hessian bias}} & \text{if } \mathfrak{p} = \ell^{\mathrm{p}}_{\cdot T}. \quad \text{(c)} \end{cases} \tag{6.10}$$

As mentioned in Example 6.1, the global maximizer of $\mathfrak{p}^{\text{adj}}_{\cdot T}$ coincides with the global maximizer of $\bar{\ell}^{\text{zse}}_{\cdot T}$. Therefore, since the estimators in (6.10)a and (6.10)b are the same, it is not surprising that the right-hand sides of (6.10)a and (6.10)b are identical. However, the sources of the $2\theta_0^2/(T-1)$ terms in (6.10)a and (6.10)b are different because $\bar{\ell}^{\text{zse}}_{iT} \neq \mathfrak{p}^{\text{adj}}_{iT}$. Indeed, since, in the Neyman–Scott model, $\bar{\ell}^{\text{zse}}_{iT}$ is {score, information}-unbiased, the $2\theta_0^2/(T-1)$ term in (6.10)a is due entirely to

[14] However, the Neyman–Scott model is an exception (cf. the discussion in Example 6.3).

[15] Except in special cases such as Example 6.3, even the signs of the O_{p} terms in (6.8)a, (6.8)b, and (6.8)c cannot be determined because it is not possible to sign the O_{p} terms in (6.4)a–(6.6) for general pseudolikelihoods. However, as discussed after 6.2, several simulation studies in the literature have found that FOSU pseudolikelihoods yield estimators with variances smaller than the fixed effects MLE.

the hessian bias of $\bar{\ell}^{\text{zse}}_{iT}$. In contrast, the $2\theta_0^2/(T-1)$ term in (6.10)b is due entirely to the information bias of $\mathfrak{p}^{\text{adj}}_{iT}$ because it is {score, hessian}-unbiased. The right-hand side of (6.10)c is a function of the score, information, and hessian biases because ℓ^{p}_{iT} is neither FOSU, nor FOIU, nor FOHU. Furthermore, unlike panel logit and probit, in the Neyman–Scott model the variances of the score bias corrected pseudologlikelihoods in finite samples are larger than the variance of the fixed effects MLE for each n, T. In fact, the finite sample variance of the fixed effects MLE is even smaller than that of the oracle estimator. □

The insights from (6.8)a and (6.8)b also apply to conditional (on sufficient statistics) likelihoods. They reveal that the conditional (on the covariates) mean-squared error (mse_0) of the CMLE in small samples may be comparable to – instead of being smaller than – the MSE of $\hat{\theta}_{\mathfrak{p}}$ if $\mathfrak{p}$ is FOSU.[16] This suggests that bias corrected estimators may be viable alternatives for CMLEs even in models where the latter exist. Indeed, although conditional likelihoods are score and information unbiased because they are genuine likelihoods (Remark 5.1), they are not FOHU in general.[17] Therefore, since there are no O_{p} terms due to the score and information bias, $\text{var}_0\sqrt{nT}(\hat{\theta}_{\text{CMLE}}-\theta_0)\overset{(6.8)\text{a}}{=}F_{.T}^{-1}+O_{\text{p}}(T^{-1})$, where the $O_{\text{p}}(T^{-1})$ term is due to the hessian bias alone. By (6.3)c, the bias of the CMLE is $O_{\text{p}}(1/nT)$. Hence,

$$\text{mse}_0(\hat{\theta}_{\text{CMLE}})=\frac{1}{nT}[F_{.T}^{-1}+O_{\text{p}}(\frac{1}{T})]+O_{\text{p}}(\frac{1}{n^2T^2}).$$

Similarly, by (6.3)a and (6.8)b,

$$\text{mse}_0(\hat{\theta}_{\mathfrak{p}})=\frac{1}{nT}[F_{.T}^{-1}+O_{\text{p}}(\frac{1}{T})+O_{\text{p}}(\frac{n}{T^3})+O_{\text{p}}(\frac{1}{T})]+O_{\text{p}}(\frac{1}{T^4}).$$

Therefore, $\text{mse}_0(\hat{\theta}_{\text{CMLE}})=\text{mse}_0(\hat{\theta}_{\mathfrak{p}})$ if n, T grow at the same rate.

In the decomposition for $\text{var}_0\sqrt{nT}(\hat{\theta}_{\mathfrak{p}}-\theta_0)-F_{.T}^{-1}$ in (6.8)a, (6.8)b, and (6.8)c, the hessian bias of $\mathfrak{p}$ is of smaller order than its information bias even when $\mathfrak{p}$ is FOIU. This naturally suggests searching for an additive correction to the profile loglikelihood such that the resulting pseudologlikelihood, in addition to being FOSU, is also FOIU with its hessian bias reduced by an order of magnitude. In other words, one can ask, "Does there exist a FOSU $\mathfrak{p}_{iT}$, which is simultaneously FOIU and FOHU?" Excluding the target likelihood, which is {score, information, hessian}-unbiased by definition, it is straightforward to show that the answer to this question is "No" in the Neyman–Scott model. In particular, for the pseudolikelihoods in Example 6.3, $\bar{\ell}^{\text{zse}}_{iT}$ is {score, information}-unbiased but not FOHU (Remark B.2), whereas $\mathfrak{p}^{\text{adj}}_{iT}$ is {score, hessian}-unbiased but not FOIU (Remark B.3). Remarkably, these are not isolated cases restricted to the Neyman–Scott model because the next result (proved in Appendix F.3) shows that:

Proposition 6.1. *Among all additive correction to a general panel data profile loglikelihood that make it* FOSU, *there does not exist one that also makes it* FOIU *and* FOHU.

For a FOSU pseudologlikelihood to be FOHU and FOIU, the model loglikelihood $\ell_{iT}(\theta,\alpha_i)$ has to satisfy a necessary condition. To see this condition, expand the profile likelihood about the target likelihood (SST, Supplement, Eqn. F.30) and take expectations to obtain

$$\mathbb{E}_0\ell^{\text{p}}_{iT}(\theta)=\mathbb{E}_0\ell_{iT}(\theta)+\mathcal{B}_{iT}(\theta)+O_{\text{p}}(T^{-2}),\tag{6.11}$$

where

$$\mathcal{B}_{iT}(\theta):=-\frac{\mathbb{E}_0\ell^2_{iT\alpha}(\theta)}{2\mathbb{E}_0\ell_{iT\alpha\alpha}(\theta)},\tag{6.12}$$

with $\ell_{iT\alpha}(\theta):=\ell_{iT\alpha}(\theta,\alpha^*_{iT}(\theta))$ and $\ell_{iT\alpha\alpha}(\theta):=\ell_{iT\alpha\alpha}(\theta,\alpha^*_{iT}(\theta))$, so that $\nabla_\theta\mathcal{B}_{iT}(\theta_0)$ is the first-order score bias of the profile loglikelihood. By (6.5) and (6.11),

$$\mathbb{E}_0\mathfrak{p}_{iT}(\theta)\overset{\text{Ass. A.2}}{=}\mathbb{E}_0\ell_{iT}(\theta)+\mathcal{B}_{iT}(\theta)+\mathbb{E}_0\mathcal{C}_{iT}(\theta)+O_{\text{p}}(T^{-2}).\tag{6.13}$$

Differentiating (6.13) twice with respect to θ and evaluating at θ_0, the hessian bias of $\mathfrak{p}_{iT}$ is

$$-H_{\mathfrak{p}_{iT}}-F_{iT}\overset{\text{Ass. A.2}}{=}-\nabla^2_{\theta\theta}\mathcal{B}_{iT}(\theta_0)-\nabla^2_{\theta\theta}\mathbb{E}_0\mathcal{C}_{iT}(\theta_0)+O_{\text{p}}(T^{-2}).\tag{6.14}$$

Since derivatives can be exchanged with expectations, and the target likelihood satisfies the 1st Bartlett identity, we have, by Assumption A.2, that

$$\mathfrak{p}_{iT}\text{ is FOSU}\overset{(6.13)}{\iff}\nabla_\theta\mathcal{B}_{iT}(\theta_0)+\mathbb{E}_0\nabla_\theta\mathcal{C}_{iT}(\theta_0)=O_{\text{p}}(T^{-2})\tag{6.15}$$

$$\mathfrak{p}_{iT}\text{ is FOHU}\overset{(6.14)}{\iff}\nabla^2_{\theta\theta}\mathcal{B}_{iT}(\theta_0)+\mathbb{E}_0\nabla^2_{\theta\theta}\mathcal{C}_{iT}(\theta_0)=O_{\text{p}}(T^{-2}).\tag{6.16}$$

[16] This may appear paradoxical at first sight because the CMLE, being semiparametrically efficient (Hahn, 1997), is expected to perform better in small samples than its competitors.

[17] E.g., in the Neyman–Scott model, the conditional likelihood is not FOHU because it coincides with the integrated likelihood of SST, which is not FOHU (Remark B.2).

DiCiccio et al. (1996, Eqn. 7) give an expression for the information bias corresponding to any additively adjusted unscaled (by the sample size) profile loglikelihood. It can be seen from this expression that, for each individual, the information bias of any additively adjusted scaled (by $1/T$) profile loglikelihood depends (modulo a $1/T^2$ term) on the expectation of the second derivative (with respect to the parameter of interest) of the adjustment term. They further show (cf. their Section 3.2) that the first-order information bias disappears if and only if the expectation of the second derivative of the adjustment term is equal to the second derivative of the additive correction proposed by Barndorff-Nielsen (1983), plus an $O_p(T^{-2})$ term. In other words,

$$\mathfrak{p}_{iT} \text{ is FOIU} \iff \mathbb{E}_0\nabla^2_{\theta\theta}\mathfrak{C}_{iT}(\theta_0) = \nabla^2_{\theta\theta}[\mathcal{A}_{iT1}(\theta_0) + \mathcal{A}_{iT2}(\theta_0)] + O_p(T^{-2}), \tag{6.17}$$

where, letting $\ell_{iT\alpha} := \ell_{iT\alpha}(\theta_0, \alpha_{i0})$, the Barndorff-Nielsen correction terms are

$$\mathcal{A}_{iT1}(\theta) := \frac{1}{2T}\log(-\mathbb{E}_0\ell_{iT\alpha\alpha}(\theta)) \quad \& \quad \mathcal{A}_{iT2}(\theta) := -\frac{1}{T}\log(T\mathbb{E}_0\ell_{iT\alpha}(\theta)\ell_{iT\alpha}). \tag{6.18}$$

If the pseudologlikelihood $\mathfrak{p}_{iT}$ is FOSU, i.e., $\mathfrak{C}_{iT}$ satisfies (6.15), then, by (6.16) and (6.17), a necessary condition for $\mathfrak{p}_{iT}$ to be FOHU and FOIU is that

$$\nabla^2_{\theta\theta}\mathcal{B}_{iT}(\theta_0) + \nabla^2_{\theta\theta}[\mathcal{A}_{iT1}(\theta_0) + \mathcal{A}_{iT2}(\theta_0)] = O_p(T^{-2}). \tag{6.19}$$

Note that (6.19) does not depend on the correction $\mathfrak{C}_{iT}$, but is a condition on the model loglikelihood. Therefore, to prove Proposition 6.1, we show in Appendix F that (6.19) cannot hold for general panel data likelihoods, excluding, of course, the target likelihood.

Showing that (6.19) does not hold for the Neyman–Scott model is straightforward.

Example 6.4 (*Example* 6.1 *Contd.*). In the Neyman–Scott model, $\ell_{iT\alpha}(\theta, \alpha_i) \overset{(B.1)}{=} \sum_{t=1}^{T}(Y_{it} - \alpha_i)/T\theta \implies \ell_{iT\alpha\alpha}(\theta, \alpha_i) = -1/\theta$. Hence, as $\alpha^*_{iT}(\theta) \overset{(B.20)}{=} \alpha_{i0}$ for each θ and $\ell_{iT\alpha} := \ell_{iT\alpha}(\theta_0, \alpha_{i0})$,

$$\ell_{iT\alpha}(\theta) := \ell_{iT\alpha}(\theta, \alpha^*_{iT}(\theta)) = \frac{1}{T\theta}\sum_{t=1}^{T}(Y_{it} - \alpha_{i0}) = \frac{1}{T\theta}\sum_{t=1}^{T}U_{it} \tag{6.20}$$

$$\ell_{iT\alpha\alpha}(\theta) := \ell_{iT\alpha\alpha}(\theta, \alpha^*_{iT}(\theta)) = -\frac{1}{\theta} \tag{6.21}$$

$$\mathbb{E}_0\ell^2_{iT\alpha}(\theta) = \frac{1}{T^2\theta^2}\sum_{t=1}^{T}\mathbb{E}_0U^2_{it} = \frac{\theta_0}{T\theta^2} \tag{6.22}$$

$$\mathbb{E}_0\ell_{iT\alpha}(\theta)\ell_{iT\alpha} \overset{(6.20)}{=} \mathbb{E}_0[\frac{1}{T\theta}\sum_{t=1}^{T}U_{it}\frac{1}{T\theta_0}\sum_{t=1}^{T}U_{it}] = \frac{1}{T^2\theta\theta_0}\sum_{t=1}^{T}\mathbb{E}_0U^2_{it} = \frac{1}{T\theta}. \tag{6.23}$$

Consequently,

$$\mathcal{B}_{iT}(\theta) \overset{(6.12)}{=} -\frac{\mathbb{E}_0\ell^2_{iT\alpha}(\theta)}{2\ell_{iT\alpha\alpha}(\theta)} \overset{(6.21)\,\&\,(6.22)}{=} \frac{\theta_0}{2T\theta} \implies \nabla^2_{\theta\theta}\mathcal{B}_{iT}(\theta_0) = \frac{1}{T\theta_0^2}$$

$$\mathcal{A}_{iT1}(\theta) \overset{(6.18)}{=} \frac{1}{2T}\log(-\ell_{iT\alpha\alpha}(\theta)) \overset{(6.21)}{=} \frac{1}{2T}\log(\theta^{-1}) \implies \nabla^2_{\theta\theta}\mathcal{A}_{iT1}(\theta_0) = \frac{1}{2T\theta_0^2}$$

$$\mathcal{A}_{iT2}(\theta) \overset{(6.18)}{=} -\frac{1}{T}\log(T\mathbb{E}_0\ell_{iT\alpha}(\theta)\ell_{iT\alpha}) \overset{(6.23)}{=} \frac{1}{T}\log(\theta) \implies \nabla^2_{\theta\theta}\mathcal{A}_{iT2}(\theta_0) = -\frac{1}{T\theta_0^2}.$$

It follows that

$$\nabla^2_{\theta\theta}\mathcal{B}_{iT}(\theta_0) + \nabla^2_{\theta\theta}[\mathcal{A}_{iT1}(\theta_0) + \mathcal{A}_{iT2}(\theta_0)] = \frac{1}{T\theta_0^2} + \frac{1}{2T\theta_0^2} - \frac{1}{T\theta_0^2} = \frac{1}{2T\theta_0^2} \neq O_p(T^{-2}).$$

Therefore, (6.19) does not hold in the Neyman–Scott model. □

7. Discussion

In this section, we summarize the relevance of our results for applied econometric practice, provide some motivation behind the proof of Lemma 6.1 (the main result in this paper), and discuss its extension to nonlinear dynamic models.

7.1. Relevance of our results for applied researchers

As mentioned in the introduction, the results obtained in this paper provide a theoretical explanation for the improved performance of modified likelihoods in small samples that researchers frequently discover in their simulation studies in the literature on nonlinear panel data models with fixed effects. The following observations, which applied researchers working in this area may find helpful, summarize our key findings:

(i) As emphasized in Example 6.1, it is possible for two different pseudolikelihoods to deliver identical estimators while exhibiting very different behavior in terms of their LR-based inference.

(ii) If the focus is on estimating the common parameters, then FOSU pseudolikelihoods should be preferred over the non-FOSU pseudolikelihoods, because maximizing them yields estimators that perform better in small samples in terms of both bias (cf. (6.3)a and (6.3)b) and variance (cf. (6.8)a–(6.8)c), as compared to the maximizers of the non-FOSU pseudolikelihoods. Furthermore, estimators that correct the bias of the profile likelihood score not only perform better in terms of bias, but may also have a lower finite sample variance than the fixed effects MLE. In particular, if n and T are comparable, then bias corrected estimators may be viable alternatives for CMLEs even in models where the latter exist (cf. the discussion following (6.8)a–(6.8)c and Example 6.3).

(iii) However, if the focus is on LR-based inference, then Lemma 6.1 shows that although FOSU pseudolikelihoods can be expected to perform substantially better than non-FOSU pseudolikelihoods, applied researchers can expect the most accurate results when using pseudolikelihoods that are both FOSU and FOIU. This also explains why confidence regions based on conditional (on sufficient statistics) likelihoods can have excellent coverage even in very short panels.

7.2. *Motivating the proof of Lemma 6.1*

Although the proof of Lemma 6.1 in Appendix C and F.1 is challenging because the remainder terms have a complicated structure, the basic idea behind the proof is easy to motivate. Assume $\dim(\theta_0) = 1$ to avoid the array notation for higher order derivatives with respect to θ. Then, by a Taylor expansion,

$$\begin{aligned}\mathbb{E}_0\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) &:= 2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T}(\hat{\theta}_{\mathfrak{p}}) - \mathfrak{p}_{\cdot T}(\theta_0)]\\ &= 2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T1}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)] + nT\mathbb{E}_0[\mathfrak{p}_{\cdot T2}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^2] + \mathrm{Rem}_{nT},\end{aligned}$$

where the additional subscripts in $\mathfrak{p}_{\cdot T}$ denote its derivatives with respect to θ (this notation is explained in Appendix A). The first two terms in the above expansion depend on whether $\mathfrak{p}$ is FOSU or FOIU. E.g., if $\mathfrak{p}$ is both FOSU and FOIU, then using an expansion for $\hat{\theta}_{\mathfrak{p}} - \theta_0$ and centering terms around their expectations, yields

$$\begin{aligned}2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T1}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)] &= 2nT\frac{\mathbb{E}_0\mathfrak{p}^2_{\cdot T1}(\theta_0)}{-\mathbb{E}_0\mathfrak{p}_{\cdot T2}(\theta_0)} + O_{\mathrm{p}}(\frac{n}{T^3}) = 2 + O_{\mathrm{p}}(\frac{n}{T^3}) + O_{\mathrm{p}}(\frac{1}{T^2})\\ nT\mathbb{E}_0[\mathfrak{p}_{\cdot T2}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^2] &= nT\frac{\mathbb{E}_0\mathfrak{p}^2_{\cdot T1}(\theta_0)}{\mathbb{E}_0\mathfrak{p}_{\cdot T2}(\theta_0)} + O_{\mathrm{p}}(\frac{n}{T^3}) = -1 + O_{\mathrm{p}}(\frac{n}{T^3}) + O_{\mathrm{p}}(\frac{1}{T^2})\\ \mathrm{Rem}_{nT} &= O_{\mathrm{p}}(\frac{1}{T^2}) + O_{\mathrm{p}}(\frac{1}{nT}),\end{aligned}$$

which leads to (6.2)a because n grows at least as fast as T. In contrast, if $\mathfrak{p}$ is neither FOSU nor FOIU, then

$$\begin{aligned}2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T1}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)] &= 2 + O_{\mathrm{p}}(\frac{n}{T}) + O_{\mathrm{p}}(\frac{1}{T})\\ nT\mathbb{E}_0[\mathfrak{p}_{\cdot T2}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^2] &= -1 + O_{\mathrm{p}}(\frac{n}{T}) + O_{\mathrm{p}}(\frac{1}{T})\\ \mathrm{Rem}_{nT} &= O_{\mathrm{p}}(\frac{n}{T}) + O_{\mathrm{p}}(\frac{n}{T^3}),\end{aligned}$$

which leads to (6.2)c.

The rates for the remainder terms in all these expansions depend on how the data are sampled (Assumption A.1), the well-behavedness of $\mathfrak{p}$ (Assumptions A.2, A.3, A.5, A.6, and A.7), and the rates at which n and T can grow (Assumption A.4). In particular, an examination of the details in the proof of Lemma 6.1 in Appendix F.1 reveals that it is essentially Assumptions A.6 and A.7 — which are justified in Appendix A under time-independence of outcomes (Assumption A.1(ii)) — that determine the rates for the remainder terms. This suggests that if Assumptions A.6 and A.7 hold for dependent outcomes as well, then so does Lemma 6.1. In fact, as shown in the next section, Assumptions A.6 and A.7 continue to hold if the outcomes are weakly dependent.

7.3. *Extending the results in Lemma 6.1 to nonlinear dynamic models*

In a nonlinear dynamic AR(p) panel data model, where $p \in \mathbb{N}$ is known and $p < T$, the period t outcome Y_{it} is modeled as a nonlinear function of the p predetermined outcomes $(Y_{i,t-p}, \ldots, Y_{i,t-1})$, additional covariates, and the fixed effects. Let $I_{i0} := (Y_{i,1-p}, Y_{i,2-p}, \ldots, Y_{i,-1}, Y_{i0})$ denote the vector of pre-sample outcomes, and $I_{i,t-1} := (I_{i0}, Y_{i1}, \ldots, Y_{i,t-1})$, $t \geq 2$.

The density of the sample outcomes, conditional on all the explanatory variables, is defined to be

$$f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},I_{i0},\alpha_{i0};\theta_0} := \begin{cases} \prod_{t=1}^{T} f_{Y_{it}|I_{i,t-1},X_{i1},\ldots,X_{it},\alpha_{i0};\theta_0} & \text{if the additional covariates are treated} \\ & \text{as being predetermined in each period} \\ \prod_{t=1}^{T} f_{Y_{it}|I_{i,t-1},X_{i1},\ldots,X_{iT},\alpha_{i0};\theta_0} & \text{if the additional covariates are treated} \\ & \text{as being strictly exogenous in each period,} \end{cases} \tag{7.1}$$

where it is assumed that the variables in I_{i0} are observed. In this setting, the pseudologlikelihood $\mathfrak{p}_{iT}(\theta)$ is obtained from $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},I_{i0},\alpha_i;\theta}$ after eliminating the fixed effects, and $(\mathbb{E}_0, \mathrm{cov}_0)$ denote the (expectation, covariance) with respect to $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},I_{i0},\alpha_{i0};\theta_0}$.

Conditional on $((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$, the dependence in the outcomes can be controlled by imposing a mixing condition as in Prakasa Rao (2009, Definition 4). Henceforth, let $\sigma((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$ denote the sigma-algebra generated by $((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$.

Definition 7.1 (*Conditional Strong Mixing, Prakasa Rao, 2009*)**.** For each $i \in \mathbb{N}$, the sequence of random variables $(Y_{it})_{t\in\mathbb{N}}$ is said to be conditionally $\mathfrak{a}_i$-mixing given $\mathcal{F}_i := \sigma((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$ if, for each $m \in \mathbb{N}$, the nonnegative random variable

$$\mathfrak{a}_i(m) := \sup_{t\in\mathbb{N}} \sup_{\substack{A\in\sigma(Y_{i1},\ldots,Y_{it}) \\ B\in\sigma(Y_{i,t+m},Y_{i,t+m+1},\ldots)}} |\Pr(A\cap B \mid \mathcal{F}_i) - \Pr(A \mid \mathcal{F}_i)\Pr(B \mid \mathcal{F}_i)| \xrightarrow[m\to\infty]{\text{w.p.1}} 0. \quad \square$$

Sufficient conditions under which AR(p)-processes are conditionally $\mathfrak{a}_i$-mixing can be derived using the approach of Andrews (1983), de Jong and Woutersen (2011), and Hahn and Kuersteiner (2011, Remarks 1–3) (cf. the discussion after Assumption A.8). E.g., following Andrews (1983) (Remark 3, p. 13), the outcomes from a dynamic Neyman–Scott model investigated in SST (Section 9.2) are conditionally $\mathfrak{a}_i$-mixing.

Under Assumption A.8, which is motivated by Hahn and Kuersteiner (2011) (Condition 3), Lemma A.1, proved in Appendix D, shows that Assumptions A.6 and A.7 continue to hold when the outcomes are conditionally $\mathfrak{a}_i$-mixing. Therefore, although a rigorous proof is beyond the scope of our paper, the analysis here suggests that, under Assumptions A.2–A.5 and A.8, Lemma 6.1 holds for nonlinear dynamic AR(p) panel data models as well. Hence, the insights from Lemma 6.1 are also relevant to the dynamic Neyman–Scott model.

8. Conclusion

We have provided a theoretical explanation behind why reducing information bias can improve the performance of likelihood based estimators and confidence regions in small samples, and why it matters more for inference than for estimation. This helps explain several simulation findings in the panel data literature. In this paper, we do not consider higher order corrections to the profile loglikelihood, even though such corrections can potentially lead to further improvements in Lemma 6.1. E.g., we conjecture that the 2nd order corrections in Dhaene and Sun (2021) and Schumann (2022) can reduce the $O_{\mathrm{p}}(n/T^3)$ term in (6.2)b to an $O_{\mathrm{p}}(n/T^5)$ term. A referee has also suggested the possibility of developing a general framework incorporating additive and non-additive corrections to the profile loglikelihood that allows for simultaneous {score, information, hessian}-bias reduction. Moreover, since time-specific parameters and dependence across i are ruled out by our sampling assumption, it would also be interesting to know the extent to which the results in this paper continue to hold in the presence of time effects and cross-sectional dependence. We leave the investigation of these issues as topics for future research.

References cited in the supplement

The following reference is cited in the supplementary material: Willink (2003).

Acknowledgments

We thank the editor Xiaohong Chen, an associate editor, and three anonymous referees, for comments that helped improve this paper. We also thank the seminar participants at the 2020 World Congress of the Econometric Society in Milan, and the 2021 Winter meeting of the Econometric Society in Chicago, for helpful suggestions and conversations.

Appendix A. Assumptions

The following assumptions are used to prove Lemma 6.1, Lemma 6.2, (6.6), and Proposition 6.1 in Appendix C and F–H. Results for the Neyman–Scott model in Appendix B and E only use Assumption A.1 and properties of Gaussian likelihoods.

Assumption A.1. (i) For each T, $(\mathcal{Y}_{1T}, \mathcal{X}_{1T}, \alpha_{10}), \dots, (\mathcal{Y}_{nT}, \mathcal{X}_{nT}, \alpha_{n0})$ are i.i.d.; (ii) For each i, conditional on $(\mathcal{X}_{iT}, \alpha_{i0})$, the outcomes $Y_{i1}, \dots, Y_{iT}$ are independent; (iii) For each i, there are no time-varying parameters or time-trends in $f_{\mathcal{Y}_{iT}|\mathcal{X}_{iT},\alpha_{i0};\theta_0}$.

(i) stipulates that observations across i are independently and identically distributed, which is typical in microeconometric applications. For each individual, (ii) imposes conditional independence within the outcomes, hereafter referred to as "time-independence", which rules out lagged outcomes as explanatory variables (cf. Assumption A.8 for dependent outcomes). This assumption greatly simplifies the algebra required for our results. Moreover, as this assumption is also used in SST, it allows us to directly use their results. Since the approximations we make are justified as $n, T \to \infty$, including an infinite number of time-specific parameters (i.e., time effects), in addition to an infinite number of fixed effects, will significantly increase the technical complexity of the proofs (cf., e.g., Fernández-Val and Weidner, 2016) while obscuring the main objective and contribution of our paper, namely, to obtain theoretical insights helpful in explaining several simulation findings in the panel data literature. Therefore, (iii), which follows if, for each i, conditional on α_{i0}, the process $(Y_{it}, X_{it})_{t\in\mathbb{N}}$ is strictly stationary, is maintained for mathematical tractability.

The next high-level assumption allows us to deal with the remainder terms in various expansions. Assumptions A.1 and A.2 can be justified following the discussion in SST (Assumptions 2.3 and C.9).

Assumption A.2. The rate at which a remainder term approaches zero (as $T \to \infty$) holds when averaged over $i = 1, \dots, n$. Taking the derivative or expectation of a remainder term does not alter the rate.

The following is assumed about θ_0 and $\hat{\theta}_{\mathfrak{p}} := \operatorname{argmax}_{\theta\in\Theta} \mathfrak{p}_{\cdot T}(\theta)$:

Assumption A.3. The common parameter θ_0 is identified as the well-separated global maximum of the limit, as $n, T \to \infty$, of the expected target loglikelihood. The estimator $\hat{\theta}_{\mathfrak{p}}$ is consistent for θ_0 as $n, T \to \infty$. Furthermore, as $n, T \to \infty$,

$$\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0) = \begin{cases} O_{\mathrm{p}}(1) + O_{\mathrm{p}}(\frac{1}{\sqrt{T}}) + O_{\mathrm{p}}(\sqrt{\frac{n}{T^3}}) & \textit{if } \mathfrak{p} \textit{ is } \text{FOSU} & \text{(a)} \\ O_{\mathrm{p}}(1) + O_{\mathrm{p}}(\sqrt{\frac{n}{T}}) & \textit{if } \mathfrak{p} \textit{ is not } \text{FOSU} & \text{(b)} \\ O_{\mathrm{p}}(1) & \textit{if } \mathfrak{p} \textit{ is the target loglikelihood.} & \text{(c)} \end{cases} \tag{A.1}$$

The assumption that θ_0 is the well-separated global maximum of the limiting expected target loglikelihood is also maintained in SST (cf. their Assumption C.5(vii) for a precise statement). Consistency of $\hat{\theta}_{\mathfrak{p}}$, which is necessary for A.3, A.3, and A.3, can be shown under standard regularity conditions; cf., e.g., SST (Theorem 8.1) for the proof of consistency of the MILE. A.3 holds for the MILE by SST (Eqn. G.7 and G.9). A.3 holds for the fixed effects MLE by SST (Eqn. C.7). The $O_{\mathrm{p}}(\sqrt{n/T^3})$ term in A.3 is due to the fact that $\mathfrak{p}$ is FOSU, and the $O_{\mathrm{p}}(T^{-1/2})$ term in A.3 arises when $\mathfrak{p}$ is approximated by the target likelihood in order to show that $\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ is asymptotically linear. If $\mathfrak{p}$ is not FOSU, then the $O_{\mathrm{p}}(\sqrt{n/T^3})$ term in A.3 becomes the $O_{\mathrm{p}}(\sqrt{n/T})$ term in A.3. Since the target likelihood is score unbiased, A.3 does not contain any $O_{\mathrm{p}}(\sqrt{n/T^3})$ or $O_{\mathrm{p}}(T^{-1/2})$ terms. Score unbiasedness of the target likelihood implies that if θ_0 also happens, for each T, to be the unique well-separated zero of the moment condition $\mathbb{E}\nabla_\theta \ell_{iT}(\theta_0) = 0$, then the oracle estimator $\hat{\theta}^*$ is consistent for θ_0 as $n \to \infty$ and T is held fixed.

The following assumption restricts the rates at which n and T can grow.

Assumption A.4. (i) n grows at least as fast as T, i.e., $\lim_{n,T\to\infty} n/T \in (0, \infty]$; (ii) T^3 grows faster than n, i.e., $\lim_{n,T\to\infty} n/T^3 = 0$.

(i) emphasizes that we are modeling short panels by restricting T to grow no faster than n: If $\lim_{n,T\to\infty} n/T \in (0, \infty)$, then n grows at the same rate as T, whereas $\lim_{n,T\to\infty} n/T = \infty$ implies that n grows faster than T; hence, if $\lim_{n,T\to\infty} n/T \in (1, \infty]$, then n is eventually (much) larger than T. Under this assumption, the $O_{\mathrm{p}}(n/T)$ terms in 6.1, 6.1, and A.3, do not vanish as $n, T \to \infty$. This ensures that if $\mathfrak{p}_{iT}$ is not FOSU, then we do not simply assume away the resulting bias in $\mathbb{E}_0\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0)$ and $\sqrt{nT}(\hat{\theta}_{\mathfrak{p}} - \theta_0)$ by making T grow faster than n. (ii) ensures that $\mathbb{E}_0\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0)$ in 6.1 and 6.1 converges to $\dim(\theta_0)$ as $n, T \to \infty$. SST (Theorem 8.2) show that the MILE is asymptotically normal if $n/T^3 \to 0$. (i) and (ii) are used repeatedly to simplify the expressions for the remainder terms.

If $\mathfrak{p}$ is FOSU, then, with $\mathfrak{r}_{nT} := 1 + T^{-1/2} + \sqrt{n/T^3}$,[18]

$$\hat{\theta}_{\mathfrak{p}} - \theta_0 \overset{\text{A.3}}{=} \underbrace{O_{\mathrm{p}}(\frac{\mathfrak{r}_{nT}}{\sqrt{nT}})}_{\text{FOSU}} \overset{\text{Ass. A.4(ii)}}{=} \underbrace{O_{\mathrm{p}}(\frac{1}{\sqrt{nT}})}_{\text{FOSU}}, \tag{A.2}$$

[18] Henceforth, the label FOSU is used to tag terms that depend on the first-order score unbiasedness of $\mathfrak{p}_{iT}$. Similarly, the label FOIU tags terms that depend on the first-order information unbiasedness of $\mathfrak{p}_{iT}$.

because $\mathfrak{r}_{nT} \overset{\text{Ass. A.4(ii)}}{=} O(1)$. If $\mathfrak{p}$ is not FOSU, then, with $\mathfrak{s}_{nT} := 1 + \sqrt{n/T}$,

$$\hat{\theta}_{\mathfrak{p}} - \theta_0 \overset{\text{A.3}}{=} O_{\mathrm{p}}(\frac{\mathfrak{s}_{nT}}{\sqrt{nT}}) = O_{\mathrm{p}}(\frac{1}{\sqrt{nT}} + \frac{1}{T}) = O_{\mathrm{p}}(\frac{1}{T}[1 + \frac{1}{\sqrt{n/T}}]) \overset{\text{Ass. A.4(i)}}{=} O_{\mathrm{p}}(\frac{1}{T}). \tag{A.3}$$

Eqs. (A.2) and (A.3) are used to bound terms in the proof of Lemma 6.1.

Additional notation. We need some additional notation for the subsequent assumptions, which hold for all the pseudolikelihoods that we consider, as well as for the target likelihood. Henceforth, denote derivatives with respect to θ by subscripts (these definitions make sense if $\dim(\theta_0) = 1$, as is assumed in Appendix C and F–H for notational simplicity). Specifically, for $k \in \{1, 2, 3, 4\}$, let $\mathfrak{p}_{\cdot Tk}(\theta) := n^{-1} \sum_{i=1}^{n} \mathfrak{p}_{iTk}(\theta)$ with $\mathfrak{p}_{iTk}(\theta) := \partial_\theta^k \mathfrak{p}_{iT}(\theta)$. In addition, letting $\lambda_{iT}(\theta) := \mathbb{E}_0 \mathfrak{p}_{iT}(\theta)$, $\lambda_{iTk}(\theta) := \mathbb{E}_0 \mathfrak{p}_{iTk}(\theta)$, $\lambda_{\cdot T}(\theta) := \mathbb{E}_0 \mathfrak{p}_{\cdot T}(\theta)$, and $\lambda_{\cdot Tk}(\theta) := \mathbb{E}_0 \mathfrak{p}_{\cdot Tk}(\theta)$, define the centered versions of $\mathfrak{p}_{iT}(\theta)$, $\mathfrak{p}_{iTk}(\theta)$, $\mathfrak{p}_{\cdot T}(\theta)$, and $\mathfrak{p}_{\cdot Tk}(\theta)$, to be $l_{iT}(\theta) := \mathfrak{p}_{iT}(\theta) - \lambda_{iT}(\theta)$, $l_{iTk}(\theta) := \mathfrak{p}_{iTk}(\theta) - \lambda_{iTk}(\theta)$, $l_{\cdot T}(\theta) := \mathfrak{p}_{\cdot T}(\theta) - \lambda_{\cdot T}(\theta)$, and $l_{\cdot Tk}(\theta) := \mathfrak{p}_{\cdot Tk}(\theta) - \lambda_{\cdot Tk}(\theta)$, respectively, so that, by construction, $\mathbb{E}_0 l_{iT}(\theta) = \mathbb{E}_0 l_{iTk}(\theta) = \mathbb{E}_0 l_{\cdot T}(\theta) = \mathbb{E}_0 l_{\cdot Tk}(\theta) = 0$.[19]

The next assumption, which can be justified under time-independence by bounding the individual summands, is used to bound terms in the proofs of the auxiliary results required to prove 6.1.

Assumption A.5. The following are bounded in probability, i.e., $O_{\mathrm{p}}(1)$, as $n, T \to \infty$: (i) $\mathfrak{p}_{\cdot T1}(\theta_0)$ and $\mathfrak{p}_{\cdot T3}(\theta_0)$; (ii) $\sup_{\theta \in \Theta} |\mathfrak{p}_{\cdot T4}(\theta)|$ and $\sup_{\theta \in \Theta} |\mathfrak{p}_{\cdot T5}(\theta)|$; (iii) $\lambda_{\cdot T2}^{-1}(\theta_0)$, $\lambda_{\cdot T3}(\theta_0)$, and $\lambda_{\cdot T4}(\theta_0)$.

The next assumption is used to prove the results in Appendix F.1.

Assumption A.6. As $n, T \to \infty$,

$$\mathbb{E}_0[l_{\cdot Tk_1}(\theta_0) l_{\cdot Tk_2}(\theta_0) l_{\cdot Tk_3}(\theta_0)] = O_{\mathrm{p}}(\frac{1}{n^2 T^2}), \qquad k_1, k_2, k_3 \in \{1, 2, 3, 4\}. \tag{A.4}$$

If $\mathfrak{p}$ does not depend on a preliminary estimator, e.g., the target likelihood, the profile likelihood, or the integrated likelihood of Lancaster, then (A.4) holds under time-independence. If $\mathfrak{p}$ depends on a preliminary estimator, e.g., the integrated likelihoods of Arellano–Bonhomme and SST, then (A.4) can be justified under time-independence and additional regularity conditions. E.g., in Appendix H.2 we show that the integrated likelihood of SST, which requires the fixed effects MLE for its construction, satisfies (A.4) under the assumptions in Appendix H.1.

The following assumption, which can also be justified under time-independence and additional regularity conditions, stipulates that the variance of the derivatives of $\mathfrak{p}_{\cdot T}$ is inversely proportional to the size of the panel data set. It is used to prove the results in Appendix F.1.

Assumption A.7. As $n, T \to \infty$,

$$\mathbb{E}_0 l^2_{\cdot Tk}(\theta_0) = \mathrm{var}_0\, \mathfrak{p}_{\cdot Tk}(\theta_0) = O_{\mathrm{p}}(\frac{1}{nT}), \qquad k \in \{1, 2, 3, 4\}. \tag{A.5}$$

It is shown in Appendix H.2 that the integrated likelihood of SST satisfies (A.5) under time-independence and additional assumptions in Appendix H.1. Since $\mathbb{E}_0 l_{\cdot Tk}(\theta_0) = 0$ by construction, an implication of (A.5) used in the proofs is that

$$l_{\cdot Tk}(\theta_0) = O_{\mathrm{p}}(\frac{1}{\sqrt{nT}}), \qquad k \in \{1, \dots, 4\}. \tag{A.6}$$

Remark A.1. Assumptions A.6 and A.7, justified earlier under time-independence, also hold when the outcomes are weakly dependent provided Assumption A.1 is suitably strengthened. The following assumption, motivated by Hahn and Kuersteiner (2011) (Condition 3), employs the notion of conditional $\mathfrak{a}_i$-mixing given in Definition 7.1; the centered derivative $l_{itk}(\theta_0)$ in (iv) is now a functional of the pseudologlikelihood resulting from $f_{\mathfrak{y}_{iT} | \mathfrak{x}_{iT}, I_{i0}, \alpha_{i0}; \theta_0}$ defined in (7.1).

Assumption A.8 (*Dependent Outcomes*)**.** (i) The random variables $((Y_{it}, X_{it})_{t \in \mathbb{N}}, I_{i0}, \alpha_{i0})$ are i.i.d. for $i \in \mathbb{N}$; (ii) For each $i \in \mathbb{N}$, the distribution of $((Y_{it}, X_{it})_{t \in \mathbb{N}}, I_{i0}, \alpha_{i0})$ is strictly stationary; (iii) The sequence of random variables $(Y_{it})_{t \in \mathbb{N}}$ is conditionally $\mathfrak{a}_i$-mixing given $\sigma((X_{it})_{t \in \mathbb{N}}, I_{i0}, \alpha_{i0})$. There exists $a \in (0, 1)$ such that, for all $m \in \mathbb{N}$, $\sup_{i \in \mathbb{N}} \mathfrak{a}_i(m) \overset{\text{w.p.1}}{\lesssim} a^m$;[20] (iv) There exists a $\sigma((X_{it})_{t \in \mathbb{N}}, I_{i0}, \alpha_{i0})$-measurable function $d_i \geq 0$ such that, for $k \in \{1, \dots, 4\}$, $\sup_{t \in \mathbb{N}} \mathbb{E}_0 |l_{itk}(\theta_0)|^8 \overset{\text{w.p.1}}{\leq} d_i$ with $n^{-1} \sum_{i=1}^{n} d_i = O_{\mathrm{p}}(1)$ as $n \to \infty$.

[19] To keep the notation simple, we suppress the dependence of λ_{iT}, $\lambda_{\cdot T}$, l_{iT}, $l_{\cdot T}$, and their derivatives, on $\mathfrak{p}$. In addition, keep in mind that in the definition of λ_{iT} the symbol $\mathbb{E}_0$ denotes expectation with respect to $f_{\mathfrak{y}_{iT} | \mathfrak{x}_{iT}, \alpha_{i0}; \theta_0}$, whereas in the definition of $\lambda_{\cdot T}$ the symbol $\mathbb{E}_0$ is expectation with respect to $\prod_{i=1}^{n} f_{\mathfrak{y}_{iT} | \mathfrak{x}_{iT}, \alpha_{i0}; \theta_0}$.

[20] Henceforth, the symbol $\lesssim$ indicates that the left-hand side of an inequality is bounded by a positive constant times the right-hand side, where the constant does not depend on n, T.

(i, ii) are the counterparts of Assumption A.1(i, iii) when the outcomes are correlated. As in Hahn and Kuersteiner (2011) (Condition 3(iii)), the bound on the mixing coefficient in (iii) implies that the dependence in the outcomes decays exponentially. Existence of the dominating function in (iv) ensures that (iii) and a conditional covariance inequality of Prakasa Rao (2009) (Theorem 9(ii), Eqn. 63) can be used to bound the covariance between elements of the sequence $(l_{itk}(\theta_0))_{t\in\mathbb{N}}$. Following the discussion in Hahn and Kuersteiner (2011) (Remarks 2 and 3), sufficient primitive conditions under which (iii, iv) hold for dynamic logit and probit panel data models can be derived using de Jong and Woutersen (2011) (Theorems 1 and 2) under the assumption that the additional covariates and the fixed effects have bounded support.

Lemma A.1. *Let Assumption A.8 hold and $n, T \to \infty$. Then, for $k_1, k_2, k_3 \in \{1, \dots, 4\}$,*

$$\mathbb{E}_0 l_{\cdot Tk_1}(\theta_0) l_{\cdot Tk_2}(\theta_0) = O_p(\frac{1}{nT}) \tag{A.7}$$

$$\mathbb{E}_0 l_{\cdot Tk_1}(\theta_0) l_{\cdot Tk_2}(\theta_0) l_{\cdot Tk_3}(\theta_0) = O_p(\frac{1}{n^2T^2}). \tag{A.8}$$

Lemma A.1, proved in Appendix D, shows that Assumptions A.6 and A.7 continue to hold when the outcomes are weakly dependent. □

Appendix B. Score, information, and hessian bias of the pseudolikelihoods in the neyman–scott model

In this section, we show that, in the Neyman–Scott model: (i) The profile likelihood is neither FOSU nor FOIU. (ii) The integrated likelihood of SST constructed with $\pi_i := 1$ is both score unbiased and information unbiased. (iii) The adjusted likelihood is score unbiased but not FOIU. (iv) The target likelihood in the Neyman–Scott model is a genuine likelihood, i.e., it satisfies all of the Bartlett identities.

Throughout this section, $\bar{Y}_{i\cdot} := T^{-1}\sum_{t=1}^{T} Y_{it}$, $\ddot{Y}_{it} := Y_{it} - \bar{Y}_{i\cdot}$, $\bar{U}_{i\cdot} := T^{-1}\sum_{t=1}^{T} U_{it}$, and $\ddot{U}_{it} := U_{it} - \bar{U}_{i\cdot}$. Keep in mind that $\ddot{Y}_{it} = \ddot{U}_{it}$ because $Y_{it} = \alpha_{i0} + U_{it}$. The results in this section only require Assumption A.1.

B.1. Profile likelihood

The Neyman–Scott model likelihood for individual i is $L_{iT}(\theta, \alpha_i) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\theta}} e^{-(Y_{it}-\alpha_i)^2/2\theta}$. Hence,

$$\ell_{iT}(\theta, \alpha_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{1}{2T\theta}\sum_{t=1}^{T}(Y_{it} - \alpha_i)^2, \tag{B.1}$$

which implies that $\hat{\alpha}_{iT}(\theta) = \bar{Y}_{i\cdot}$. Therefore, the profile loglikelihood for individual i is

$$\ell^{\mathrm{p}}_{iT}(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{1}{2T\theta}\sum_{t=1}^{T}\ddot{Y}^2_{it}. \tag{B.2}$$

By (B.2), and the fact that $\ddot{Y}_{it} = \ddot{U}_{it}$,

$$\nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta) = -\frac{1}{2\theta} + \frac{1}{2T\theta^2}\sum_{t=1}^{T}\ddot{U}^2_{it} \quad \& \quad \nabla^2_{\theta\theta}\ell^{\mathrm{p}}_{iT}(\theta) = \frac{1}{2\theta^2} - \frac{1}{T\theta^3}\sum_{t=1}^{T}\ddot{U}^2_{it}. \tag{B.3}$$

Since $\mathfrak{U}_{iT} := (U_{i1}, \dots, U_{iT})_{T\times 1} \mid \alpha_{i0} \overset{\mathrm{d}}{=} \mathrm{N}(0_{d\times 1}, \theta_0 I_T)$, where I_T is the $T \times T$ identity matrix,

$$\sum_{t=1}^{T}\ddot{U}^2_{it} = \mathfrak{U}'_{iT}(I_T - \frac{1}{T}1_T 1'_T)\mathfrak{U}_{iT} \mid \alpha_{i0} \overset{\mathrm{d}}{=} \theta_0 \chi^2_{T-1} \tag{B.4}$$

by Cochran's theorem, where $1_T := (1, \dots, 1)_{T\times 1}$. Therefore (cf. (E.31)),

$$\mathbb{E}_0 \sum_{t=1}^{T}\ddot{U}^2_{it} = (T-1)\theta_0 \quad \& \quad \mathbb{E}_0[\sum_{t=1}^{T}\ddot{U}^2_{it}]^2 = (T^2-1)\theta_0^2. \tag{B.5}$$

By (B.3) and (B.5),

$$\mathbb{E}_0 \nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta) = -\frac{1}{2\theta} + \frac{1}{2T\theta^2}(T-1)\theta_0 \implies \mathbb{E}_0 \nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta_0) = -\frac{1}{2\theta_0 T}. \tag{B.6}$$

Hence, the profile likelihood is not FOSU. By (B.3) and (B.5) again,

$$\mathbb{E}_0 \nabla^2_{\theta\theta}\ell^{\mathrm{p}}_{iT}(\theta) = \frac{1}{2\theta^2} - \frac{1}{T\theta^3}(T-1)\theta_0 \implies \mathbb{E}_0 \nabla^2_{\theta\theta}\ell^{\mathrm{p}}_{iT}(\theta_0) = -\frac{1}{2\theta_0^2} + \frac{1}{T\theta_0^2} \tag{B.7}$$

and

$$\mathbb{E}_0[\nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta)]^2 = \mathbb{E}_0[-\frac{1}{2\theta} + \frac{1}{2T\theta^2}\sum_{t=1}^{T} \ddot{U}_{it}^2]^2 = \frac{1}{4\theta^2} + \frac{1}{4T^2\theta^4}(T^2-1)\theta_0^2 - \frac{2}{4T\theta^3}(T-1)\theta_0,$$

which implies that

$$\mathbb{E}_0[\nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta_0)]^2 = \frac{1}{2T\theta_0^2} - \frac{1}{4T^2\theta_0^2}. \tag{B.8}$$

Hence,

$$T\mathbb{E}_0[\nabla_\theta \ell^{\mathrm{p}}_{iT}(\theta_0)]^2 + \mathbb{E}_0\nabla^2_{\theta\theta}\ell^{\mathrm{p}}_{iT}(\theta_0) = \frac{3}{4\theta_0^2 T}. \tag{B.9}$$

Therefore, the profile likelihood is also not FOIU.

Remark B.1. By (B.6) and (B.9), the profile likelihood in the Neyman–Scott model is neither FOSU nor FOIU. It is also not FOHU because $F_{iT} \overset{\text{Remark B.4}}{=} (2\theta_0^2)^{-1}$ and, by (B.7), $-\mathbb{E}_0\nabla^2_{\theta\theta}\ell^{\mathrm{p}}_{iT}(\theta_0) - F_{iT} = (-T\theta_0^2)^{-1}$. □

B.2. Integrated likelihood of SST

In the Neyman–Scott model, the ZSE transformation and its inverse are the identity maps from $\mathbb{R} \to \mathbb{R}$ (SST, Example 9.1). Hence, the individual integrated likelihood of SST constructed with $\pi_i := 1$ is given by

$$\bar{L}^{\text{zse}}_{iT}(\theta) = \int_{\mathbb{R}} L_{iT}(\theta,\phi)\,d\phi = (2\pi)^{-(T-1)/2}T^{-1/2}\theta^{-(T-1)/2}e^{-\sum_{t=1}^{T}\ddot{Y}_{it}^2/2\theta},$$

Therefore, the integrated loglikelihood of SST for individual i is

$$\bar{\ell}^{\text{zse}}_{iT}(\theta) = -\frac{T-1}{2T}\log(2\pi) - \frac{1}{2T}\log(T) - \frac{T-1}{2T}\log(\theta) - \frac{1}{2T\theta}\sum_{t=1}^{T}\ddot{Y}_{it}^2. \tag{B.10}$$

By (B.10), and the fact that $\ddot{Y}_{it} = \ddot{U}_{it}$,

$$\nabla_\theta\bar{\ell}^{\text{zse}}_{iT}(\theta) = -\frac{T-1}{2T\theta} + \frac{1}{2T\theta^2}\sum_{t=1}^{T}\ddot{U}_{it}^2 \quad \& \quad \nabla^2_{\theta\theta}\bar{\ell}^{\text{zse}}_{iT}(\theta) = \frac{T-1}{2T\theta^2} - \frac{1}{T\theta^3}\sum_{t=1}^{T}\ddot{U}_{it}^2. \tag{B.11}$$

By (B.5) and (B.11),

$$\mathbb{E}_0\nabla_\theta\bar{\ell}^{\text{zse}}_{iT}(\theta) = -\frac{T-1}{2T\theta} + \frac{1}{2T\theta^2}(T-1)\theta_0.$$

Consequently, it is immediate that

$$\mathbb{E}_0\nabla_\theta\bar{\ell}^{\text{zse}}_{iT}(\theta_0) = 0, \tag{B.12}$$

i.e., the integrated likelihood of SST is score unbiased. Moreover, by (B.5) and (B.11),

$$\begin{aligned}\mathbb{E}_0\nabla^2_{\theta\theta}\bar{\ell}^{\text{zse}}_{iT}(\theta) &= \frac{T-1}{2T\theta^2} - \frac{1}{T\theta^3}(T-1)\theta_0\\ \mathbb{E}_0[\nabla_\theta\bar{\ell}^{\text{zse}}_{iT}(\theta)]^2 &= \frac{(T-1)^2}{4T^2\theta^2} + \frac{1}{4T^2\theta^4}(T^2-1)\theta_0^2 - \frac{2(T-1)}{4T^2\theta^3}(T-1)\theta_0.\end{aligned} \tag{B.13}$$

Hence, it is straightforward to verify that

$$T\mathbb{E}_0[\nabla_\theta\bar{\ell}^{\text{zse}}_{iT}(\theta_0)]^2 + \mathbb{E}_0\nabla^2_{\theta\theta}\bar{\ell}^{\text{zse}}_{iT}(\theta_0) = 0, \tag{B.14}$$

i.e., the integrated likelihood of SST is also information unbiased.

Remark B.2. By (B.12) and (B.14), the integrated likelihood of SST in the Neyman–Scott model is {score, information}-unbiased. However, $\bar{\ell}^{\text{zse}}_{iT}$ is not FOHU because $F_{iT} \overset{\text{Remark B.4}}{=} (2\theta_0^2)^{-1}$ and $-\mathbb{E}_0\nabla^2_{\theta\theta}\bar{\ell}^{\text{zse}}_{iT}(\theta_0) - F_{iT} \overset{(B.13)}{=} (-2T\theta_0^2)^{-1}$. □

B.3. Explicitly adjusted likelihood

The explicitly adjusted loglikelihood for individual i is

$$\mathfrak{p}^{\text{adj}}_{iT}(\theta) := \ell^{\mathrm{p}}_{iT}(\theta) + \mathfrak{c}_{iT}(\theta) \overset{(B.2)}{=} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{1}{2T\theta}\sum_{t=1}^{T}\ddot{Y}_{it}^2 + \mathfrak{c}_{iT}(\theta), \tag{B.15}$$

where

$$\mathfrak{C}_{iT}(\theta) := -\frac{T^{-1}\sum_{t=1}^{T}\ddot{Y}_{it}^{2}}{2(T-1)\theta}. \tag{B.16}$$

By (B.15), and the fact that $\ddot{Y}_{it} = \ddot{U}_{it}$,

$$\nabla_{\theta}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta) = \nabla_{\theta}\ell_{iT}^{\mathrm{p}}(\theta) + \frac{T^{-1}\sum_{t=1}^{T}\ddot{U}_{it}^{2}}{2(T-1)\theta^{2}} \quad \& \quad \nabla_{\theta\theta}^{2}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta) = \nabla_{\theta\theta}^{2}\ell_{iT}^{\mathrm{p}}(\theta) - \frac{T^{-1}\sum_{t=1}^{T}\ddot{U}_{it}^{2}}{(T-1)\theta^{3}}.$$

Consequently, by (B.5) and (B.6),

$$\mathbb{E}_0\nabla_{\theta}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta_0) = \mathbb{E}_0\nabla_{\theta}\ell_{iT}^{\mathrm{p}}(\theta_0) + \frac{\mathbb{E}_0 T^{-1}\sum_{t=1}^{T}\ddot{U}_{it}^{2}}{2(T-1)\theta_0^{2}} = -\frac{1}{2\theta_0 T} + \frac{1}{2\theta_0 T} = 0, \tag{B.17}$$

i.e., $\mathfrak{p}_{iT}^{\mathrm{adj}}$ is score unbiased. Next, by (B.5) again,

$$\mathbb{E}_0\nabla_{\theta\theta}^{2}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta_0) = \mathbb{E}_0\nabla_{\theta\theta}^{2}\ell_{iT}^{\mathrm{p}}(\theta_0) - \frac{\mathbb{E}_0 T^{-1}\sum_{t=1}^{T}\ddot{U}_{it}^{2}}{(T-1)\theta_0^{3}} \overset{(\mathrm{B.7})}{=} -\frac{1}{2\theta_0^{2}} + \frac{1}{T\theta_0^{2}} - \frac{1}{T\theta_0^{2}} = -\frac{1}{2\theta_0^{2}}. \tag{B.18}$$

Moreover,

$$\begin{aligned}\mathbb{E}_0[\nabla_{\theta}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta_0)]^{2} &= \mathbb{E}_0[\nabla_{\theta}\ell_{iT}^{\mathrm{p}}(\theta_0)]^{2} + \frac{\mathbb{E}_0[T^{-1}\sum_{t=1}^{T}\ddot{U}_{it}^{2}]^{2}}{4(T-1)^{2}\theta_0^{4}} + \frac{\mathbb{E}_0[\nabla_{\theta}\ell_{iT}^{\mathrm{p}}(\theta_0)T^{-1}\sum_{t=1}^{T}\ddot{U}_{it}^{2}]}{(T-1)\theta_0^{2}} \\ &\overset{(\mathrm{B.8})\ \&\ (\mathrm{B.5})}{=} \frac{1}{2T\theta_0^{2}} - \frac{1}{4T^{2}\theta_0^{2}} + \frac{T+1}{4T^{2}(T-1)\theta_0^{2}} + \frac{\mathbb{E}_0[\nabla_{\theta}\ell_{iT}^{\mathrm{p}}(\theta_0)\sum_{t=1}^{T}\ddot{U}_{it}^{2}]}{T(T-1)\theta_0^{2}}.\end{aligned}$$

By (B.3) and (B.5),

$$\mathbb{E}_0[\nabla_{\theta}\ell_{iT}^{\mathrm{p}}(\theta_0)\sum_{t=1}^{T}\ddot{U}_{it}^{2}] = -\frac{\mathbb{E}_0\sum_{t=1}^{T}\ddot{U}_{it}^{2}}{2\theta_0} + \frac{1}{2T\theta_0^{2}}\mathbb{E}_0[\sum_{t=1}^{T}\ddot{U}_{it}^{2}]^{2} = -\frac{T-1}{2} + \frac{T^{2}-1}{2T} = \frac{T-1}{2T}.$$

It follows that

$$\mathbb{E}_0[\nabla_{\theta}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta_0)]^{2} = \frac{1}{2T\theta_0^{2}} - \frac{1}{4T^{2}\theta_0^{2}} + \frac{T+1}{4T^{2}(T-1)\theta_0^{2}} + \frac{1}{2T^{2}\theta_0^{2}} = \frac{1}{2T\theta_0^{2}} + \frac{1}{2T(T-1)\theta_0^{2}}.$$

Therefore,

$$T\mathbb{E}_0[\nabla_{\theta}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta_0)]^{2} + \mathbb{E}_0\nabla_{\theta\theta}^{2}\mathfrak{p}_{iT}^{\mathrm{adj}}(\theta_0) = \frac{1}{2\theta_0^{2}} + \frac{1}{2(T-1)\theta_0^{2}} - \frac{1}{2\theta_0^{2}} = \frac{1}{2(T-1)\theta_0^{2}}. \tag{B.19}$$

Consequently, $\mathfrak{p}_{iT}^{\mathrm{adj}}$ is not FOIU.

Remark B.3. $\mathfrak{p}_{iT}^{\mathrm{adj}}$ is {score, hessian}-unbiased by (B.17) and (B.18) because $F_{iT} \overset{\text{Remark B.4}}{=} (2\theta_0^{2})^{-1}$. However, (B.19) shows that $\mathfrak{p}_{iT}^{\mathrm{adj}}$ is not FOIU. □

B.4. Target likelihood

Recall that $\alpha_{iT}^{*}(\theta)$ solves $\mathbb{E}_0\ell_{iT\alpha}(\theta, \alpha_{iT}^{*}(\theta)) = 0$. But $\mathbb{E}_0\ell_{iT\alpha}(\theta, \alpha_i) \overset{(\mathrm{B.1})}{=} \sum_{t=1}^{T}(\mathbb{E}_0 Y_{it} - \alpha_i)/T\theta = (\alpha_{i0} - \alpha_i)/\theta$. Hence,

$$\alpha_{iT}^{*}(\theta) = \alpha_{i0} \quad \text{for each } \theta. \tag{B.20}$$

Therefore, in the Neyman–Scott model, the target loglikelihood for individual i is

$$\ell_{iT}(\theta) := \ell_{iT}(\theta, \alpha_{iT}^{*}(\theta)) \overset{(\mathrm{B.1})}{=} -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\theta) - \frac{1}{2T\theta}\sum_{t=1}^{T}U_{it}^{2}. \tag{B.21}$$

Since $\ell_{iT}(\theta) \overset{(\mathrm{B.21})}{=} T^{-1}\log\prod_{t=1}^{T}\mathrm{pdf}_{\mathrm{N}(0,\theta)}(U_{it})$, the target likelihood $\prod_{t=1}^{T}\mathrm{pdf}_{\mathrm{N}(0,\theta)}(U_{it})$ satisfies all of the Bartlett identities because $\int_{\mathbb{R}^T}\prod_{t=1}^{T}\mathrm{pdf}_{\mathrm{N}(0,\theta)}(u_{it})\,du_{i1}\ldots u_{iT} = 1$ for each θ.

Remark B.4. The target likelihood is {score, information }-unbiased as it satisfies all of the Bartlett identities, and it is hessian unbiased by definition. Furthermore, $\mathbb{E}_0\nabla_{\theta\theta}^{2}\ell_{iT}(\theta) \overset{(\mathrm{B.21})}{=} \frac{1}{2\theta^{2}} - \frac{\theta_0}{\theta^{3}}$. Therefore, in the Neyman–Scott model, $F_{iT} := -\mathbb{E}_0\nabla_{\theta\theta}^{2}\ell_{iT}(\theta_0) = (2\theta_0^{2})^{-1} \implies F_{\cdot T} = (2\theta_0^{2})^{-1}$, which confirms (6.9). □

Appendix C. Proof of Lemma 6.1

To keep the notation for higher order derivatives with respect to θ from becoming unwieldy, assume henceforth that $\dim(\theta_0) = 1$. The results in Lemma 6.1 and the remainder of the paper hold even if $\dim(\theta_0) > 1$ because the rates for the remainder terms hold coordinatewise, although their proofs become very tedious due to the cumbersome notation.

Proof (***Proof of*** **6.1**)**.** Recall that $\mathbb{E}_0\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) := 2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T}(\hat{\theta}_{\mathfrak{p}}) - \mathfrak{p}_{\cdot T}(\theta_0)]$. By Assumption A.3, $\hat{\theta}_{\mathfrak{p}}$ is consistent for θ_0 as $n, T \to \infty$. Hence, expanding $\mathfrak{p}_{\cdot T}(\hat{\theta}_{\mathfrak{p}})$ about θ_0 and integrating with respect to $\prod_{i=1}^{n} f_{\mathfrak{y}_{iT}|\mathfrak{x}_{iT},\alpha_{i0};\theta_0}$, we have

$$\begin{aligned}&\mathbb{E}_0\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) = 2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T1}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)] + nT\mathbb{E}_0[\mathfrak{p}_{\cdot T2}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^2]\\ &+\frac{nT}{3}\mathbb{E}_0[\mathfrak{p}_{\cdot T3}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^3] + \frac{nT}{12}\mathbb{E}_0[\mathfrak{p}_{\cdot T4}(\bar{\theta})(\hat{\theta}_{\mathfrak{p}} - \theta_0)^4],\end{aligned} \tag{C.1}$$

where $\bar{\theta}$ lies between $\hat{\theta}_{\mathfrak{p}}$ and θ_0. As shown in Appendix F.1, first-order score unbiasedness and first-order information unbiasedness of $\mathfrak{p}_{iT}$ implies that, as $n, T \to \infty$,

$$2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T1}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)] = 2 + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\text{FOSU}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{FOIU}} \tag{C.2}$$

$$nT\mathbb{E}_0[\mathfrak{p}_{\cdot T2}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^2] = -1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\text{FOSU}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{FOIU}} \tag{C.3}$$

$$nT\mathbb{E}_0[\mathfrak{p}_{\cdot T3}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^3] = \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{FOSU}} \tag{C.4}$$

$$nT\mathbb{E}_0[\mathfrak{p}_{\cdot T4}(\bar{\theta})(\hat{\theta}_{\mathfrak{p}} - \theta_0)^4] = \underbrace{O_{\mathrm{p}}(\frac{1}{nT})}_{\text{FOSU}}. \tag{C.5}$$

By (C.1)–(C.5),

$$\begin{aligned}\mathbb{E}_0\mathrm{LR}^{\mathfrak{p}}_{nT}(\theta_0) &= 1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3}) + O_{\mathrm{p}}(\frac{1}{T^2}) + O_{\mathrm{p}}(\frac{1}{nT})}_{\text{FOSU}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{FOIU}}\\ &= 1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3}[1 + \frac{1}{n/T} + \frac{1}{n^2/T^2}])}_{\text{FOSU}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{FOIU}} \overset{\text{Ass.A.4(i)}}{=} 1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\text{FOSU}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T^2})}_{\text{FOIU}}. \quad \square\end{aligned} \tag{C.6}$$

Proof (***Proof of (6.2)b***)**.** Except for one change, namely, that $\mathfrak{p}_{iT}$ is no longer FOIU, (6.2)b is proved exactly as (6.2)a. The terms due to the first-order information unbiasedness of $\mathfrak{p}_{iT}$ only occur in the proofs of (C.2) and (C.3), and are clearly marked. Therefore, the desired result follows from the proof of (6.2)a by modifying these terms as explained in Footnotes 27 and 30.

Proof (***Proof of (6.2)c***)**.** If $\mathfrak{p}_{iT}$ is not FOSU, then $\lambda_{\cdot T1}(\theta_0) = O_{\mathrm{p}}(T^{-1})$ and $\hat{\theta}_{\mathfrak{p}} - \theta_0 \overset{(A.3)}{=} O_{\mathrm{p}}(T^{-1})$. Moreover, if $\mathfrak{p}_{iT}$ is not FOIU, then the right-hand side of (3.2) is $O_{\mathrm{p}}(T^{-1})$. We replicate the proof of (6.2)a, modifying it where necessary to account for these facts. Let the labels "SB" (short for score bias) and "IB" (short for information bias) tag terms depending on the score and information bias, respectively. As shown in Appendix F.1, if $\mathfrak{p}_{iT}$ is neither FOSU, nor FOIU, then as $n, T \to \infty$,

$$2nT\mathbb{E}_0[\mathfrak{p}_{\cdot T1}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)] = 2 + \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{\text{SB}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{IB}} \tag{C.7}$$

$$nT\mathbb{E}_0[\mathfrak{p}_{\cdot T2}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^2] = -1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{\text{SB}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\text{IB}} \tag{C.8}$$

$$nT\mathbb{E}_0[\mathfrak{p}_{\cdot T3}(\theta_0)(\hat{\theta}_{\mathfrak{p}} - \theta_0)^3] = \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{\text{SB}} \tag{C.9}$$

$$nT\mathbb{E}_0[\mathfrak{p}_{\cdot T4}(\bar{\theta})(\hat{\theta}_{\mathfrak{p}} - \theta_0)^4] = \underbrace{O_{\mathrm{p}}(\frac{n}{T^3})}_{\text{SB}}, \tag{C.10}$$

where $\bar{\theta}$ lies between $\hat{\theta}_{\mathrm{p}}$ and θ_0. By (C.1) and (C.7)–(C.10),

$$\mathbb{E}_0\mathrm{LR}^{\mathrm{p}}_{nT}(\theta_0) = 1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T}) + O_{\mathrm{p}}(\frac{n}{T^3})}_{\mathrm{SB}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\mathrm{IB}} = 1 + \underbrace{O_{\mathrm{p}}(\frac{n}{T})}_{\mathrm{SB}} + \underbrace{O_{\mathrm{p}}(\frac{1}{T})}_{\mathrm{IB}}. \quad \square$$

Proof (***Proof of (6.2)d***)**.** We replicate the proof of (6.2)a, modifying it where necessary to account for the fact that the target likelihood is a genuine likelihood. Recall that $\mathbb{E}_0\mathrm{LR}^{\mathrm{target}}_{nT}(\theta_0) := 2nT\mathbb{E}_0[\ell_{\cdot T}(\hat{\theta}^*) - \ell_{\cdot T}(\theta_0)]$. By Assumption A.3, $\hat{\theta}^*$ is consistent for θ_0 as $n, T \to \infty$. Hence, expanding $\ell_{\cdot T}(\hat{\theta}^*)$ about θ_0 and integrating with respect to $\prod_{i=1}^n f_{\mathscr{y}_{iT}|\mathscr{x}_{iT},\alpha_{i0};\theta_0}$, we get

$$\begin{aligned}\mathbb{E}_0\mathrm{LR}^{\mathrm{target}}_{nT}(\theta_0) &= 2nT\mathbb{E}_0[\ell_{\cdot T1}(\theta_0)(\hat{\theta}^* - \theta_0)] + nT\mathbb{E}_0[\ell_{\cdot T2}(\theta_0)(\hat{\theta}^* - \theta_0)^2] \\ &+ \frac{nT}{3}\mathbb{E}_0[\ell_{\cdot T3}(\theta_0)(\hat{\theta}^* - \theta_0)^3] + \frac{nT}{12}\mathbb{E}_0[\ell_{\cdot T4}(\bar{\theta})(\hat{\theta}^* - \theta_0)^4], \end{aligned} \tag{C.11}$$

where $\bar{\theta}$ lies between $\hat{\theta}^*$ and θ_0. As shown in Appendix F.1, since the target likelihood is a genuine likelihood, we have, as $n, T \to \infty$,

$$2nT\mathbb{E}_0[\ell_{\cdot T1}(\theta_0)(\hat{\theta}^* - \theta_0)] = 2 + O_{\mathrm{p}}(\frac{1}{nT}) \tag{C.12}$$

$$nT\mathbb{E}_0[\ell_{\cdot T2}(\theta_0)(\hat{\theta}^* - \theta_0)^2] = -1 + O_{\mathrm{p}}(\frac{1}{nT}) \tag{C.13}$$

$$nT\mathbb{E}_0[\ell_{\cdot T3}(\theta_0)(\hat{\theta}^* - \theta_0)^3] = O_{\mathrm{p}}(\frac{1}{nT}) \tag{C.14}$$

$$nT\mathbb{E}_0[\ell_{\cdot T4}(\bar{\theta})(\hat{\theta}^* - \theta_0)^4] = O_{\mathrm{p}}(\frac{1}{nT}). \tag{C.15}$$

By (C.11)–(C.15), $\mathbb{E}_0\mathrm{LR}^{\mathrm{target}}_{nT}(\theta_0) = 1 + O_{\mathrm{p}}(\frac{1}{nT})$.

Appendix D. Proof of Lemma A.1

Let $l_{\cdot Tk} := l_{\cdot Tk}(\theta_0)$ for notational convenience. Since the centered derivative

$$l_{\cdot Tk} = \frac{1}{n}\sum_{i=1}^n[\mathfrak{p}_{iTk} - \mathbb{E}_0\mathfrak{p}_{iTk}] = \frac{1}{n}\sum_{i=1}^n l_{iTk} = \frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T l_{itk}$$

is a transformation of only finitely many Y_{it}, we can use Davidson (1994, Theorem 14.1) to conclude that $l_{\cdot Tk}$ is conditionally $\mathfrak{a}_i$-mixing with $\sup_{i\in\mathbb{N}}\mathfrak{a}_i(m) \overset{\text{w.p.1}}{\lesssim} a^m$. Moreover, the same holds for the products of the centered derivatives in (A.7) and (A.8).

We now show (A.7). Observe that $\mathbb{E}_0 l_{\cdot Tk_1} l_{\cdot Tk_2} = n^{-2}\sum_{i=1}^n \mathbb{E}_0 l_{iTk_1} l_{iTk_2}$ by independence across i and the fact that $\mathbb{E}_0 l_{iTk} = 0$ for each i by construction. Now,

$$\begin{aligned}\mathbb{E}_0 l_{iTk_1} l_{iTk_2} &= \frac{1}{T^2}\sum_{t=1}^T\sum_{s=1}^T \mathbb{E}_0 l_{itk_1} l_{isk_2} \\ &= 1 \text{ term of the form } \frac{1}{T^2}\sum_{t=1}^T \mathbb{E}_0 l_{itk_1} l_{itk_2} \text{ with 1 index} \\ &\quad + 2 \text{ terms of the form } \frac{1}{T^2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}_0 l_{itk_1} l_{isk_2}, \text{ each with 2 different indices.}\end{aligned}$$

Since $0 < m \le p \implies \forall x \in \mathbb{R}, \quad |x|^m \le \mathbb{1}(|x| \le 1) + |x|^p$, we have that

$$0 < m \le 8 \implies \mathbb{E}_0|l_{itk}|^m \overset{\text{w.p.1}}{\le} 1 + \mathbb{E}_0 l_{itk}^8. \tag{D.1}$$

Hence,

$$\begin{aligned}\frac{1}{T^2}\sum_{t=1}^T \mathbb{E}_0|l_{itk_1} l_{itk_2}| &\overset{\text{AM-GM ineq.}}{\le} \frac{1}{T^2}\sum_{t=1}^T \frac{\mathbb{E}_0 l_{itk_1}^2 + \mathbb{E}_0 l_{itk_2}^2}{2} \quad \text{w.p.1} \\ &\overset{\text{(D.1)}}{\lesssim} \frac{1}{T^2}\sum_{t=1}^T(1 + \mathbb{E}_0 l_{itk_1}^8 + \mathbb{E}_0 l_{itk_2}^8) \\ &\overset{\text{(iv)}}{\lesssim} \frac{1 + d_i}{T}.\end{aligned}$$

Moreover, given $((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$, l_{itk_1} is $\sigma(Y_{i1},\dots,Y_{it})$-measurable, whereas, for $s > t$, l_{isk_2} is $\sigma(Y_{i,t+m}, Y_{i,t+m+1},\dots)$-measurable with $m := s-t$. Hence,

$$|\mathbb{E}_0 l_{itk_1} l_{isk_2}| = |\mathrm{cov}_0(l_{itk_1}, l_{i,t+m,k_2})| \overset{\text{w.p.1}}{\lesssim} (3+2d_i)\mathfrak{a}_i^{1/2}(m) \overset{\text{(iii)}}{\lesssim} (1+d_i)a^{m/2},$$

where the first inequality follows from a conditional covariance inequality for strong mixing sequences of random variables (Prakasa Rao (2009), Eqn. 63) because $\mathbb{E}_0 l_{itk}^4 \overset{\text{(D.1)}}{\lesssim} 1 + \mathbb{E}_0 l_{itk}^8 \overset{\text{(iv)}}{\lesssim} 1 + d_i$. Thus,

$$\begin{aligned}
\frac{1}{T^2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} |\mathbb{E}_0 l_{itk_1} l_{isk_2}| &\overset{\text{w.p.1}}{\lesssim} \frac{(1+d_i)}{T^2}\sum_{t=1}^{T-1}\sum_{m=1}^{T-t} b^m && (b := a^{1/2})\\
&= \frac{(1+d_i)}{T^2}\sum_{t=1}^{T-1}(\frac{1-b^{T-t}}{1-b})b\\
&\lesssim \frac{(1+d_i)}{T^2}\sum_{u=1}^{T-1}(1-b^u) && (u := T-t)\\
&= \frac{(1+d_i)}{T^2}[T-1-(\frac{1-b^{T-1}}{1-b})b]\\
&\lesssim \frac{1+d_i}{T}. && \text{(D.2)}
\end{aligned}$$

Combining these results, we have that

$$\mathbb{E}_0 l_{\cdot Tk_1} l_{\cdot Tk_2} = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_0 l_{iTk_1} l_{iTk_2} \overset{\text{w.p.1}}{\lesssim} \frac{1}{n^2}\sum_{i=1}^{n}\frac{1+d_i}{T} \overset{\text{(iv)}}{=} O_p(\frac{1}{nT}),$$

which yields (A.7).

Next, we show (A.8). Observe that $\mathbb{E}_0 l_{\cdot Tk_1} l_{\cdot Tk_2} l_{\cdot Tk_3} = n^{-3}\sum_{i=1}^{n} \mathbb{E}_0 l_{iTk_1} l_{iTk_2} l_{iTk_3}$ by independence across i and the fact that $\mathbb{E}_0 l_{iTk} = 0$ for each i by construction. Now,

$$\begin{aligned}
&\mathbb{E}_0 l_{iTk_1} l_{iTk_2} l_{iTk_3}\\
&\quad = \frac{1}{T^3}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{r=1}^{T} l_{itk_1} l_{isk_2} l_{irk_3}\\
&\quad = 1 \text{ term of the form } \frac{1}{T^3}\sum_{t=1}^{T} \mathbb{E}_0 l_{itk_1} l_{itk_2} l_{itk_3} \text{ with 1 index}\\
&\qquad + 3 \text{ terms of the form } \frac{1}{T^3}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \mathbb{E}_0 l_{itk_1} l_{itk_2} l_{isk_3}\text{, each with 2 different indices}\\
&\qquad + 1 \text{ term of the form } \frac{1}{T^3}\sum_{t=1}^{T-2}\sum_{s=t+1}^{T-1}\sum_{r=s+1}^{T} \mathbb{E}_0 l_{itk_1} l_{isk_2} l_{irk_3} \text{ with 3 different indices.}
\end{aligned}$$

Since

$$\mathbb{E}_0|l_{itk_1} l_{itk_2} l_{itk_3}| \overset{\text{AM-GM}}{\le} \frac{1}{3}\mathbb{E}_0[|l_{itk_1}|^3 + |l_{itk_2}|^3 + |l_{itk_3}|^3] \overset{\text{(D.1)}}{\lesssim} 1 + \mathbb{E}_0|l_{itk_1}|^8 + \mathbb{E}_0|l_{itk_2}|^8 + \mathbb{E}_0|l_{itk_3}|^8,$$

we have that

$$\frac{1}{T^3}\sum_{t=1}^{T} |\mathbb{E}_0 l_{itk_1} l_{itk_2} l_{itk_3}| \overset{\text{(iv)}}{\le} \frac{1+d_i}{T^2} \quad \text{w.p.1.}$$

Next, given $((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$, l_{itk_1} and l_{itk_2} are $\sigma(Y_{i1},\dots,Y_{it})$-measurable, whereas, for $s > t$, l_{isk_3} is $\sigma(Y_{i,t+m}, Y_{i,t+m+1},\dots)$-measurable with $m := s-t$. Hence, as

$$\mathbb{E}_0(l_{itk_1} l_{itk_2})^4 \overset{\text{AM-GM}}{\le} \frac{\mathbb{E}_0 l_{itk_1}^8 + \mathbb{E}_0 l_{itk_2}^8}{2} \overset{\text{(iv)}}{\le} d_i \quad \& \quad \mathbb{E}_0 l_{i,t+m,k_3}^4 \overset{\text{(D.1)}}{\le} 1 + \mathbb{E}_0 l_{i,t+m,k_3}^8 \overset{\text{(iv)}}{\le} 1 + d_i,$$

we have that

$$|\mathbb{E}_0 l_{itk_1} l_{itk_2} l_{isk_3}| = |\mathrm{cov}_0(l_{itk_1} l_{itk_2}, l_{i,t+m,k_3})| \overset{\text{w.p.1}}{\lesssim} (1+d_i)\mathfrak{a}_i^{1/2}(m) \tag{Prakasa Rao, Eqn. 63}$$

$$\overset{\text{(iii)}}{\lesssim} (1+d_i)a^{m/2}.$$

Consequently, following the argument leading to (D.2),

$$\frac{1}{T^3}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \mathbb{E}_0 l_{itk_1} l_{itk_2} l_{isk_3} = \frac{1}{T^3}\sum_{t=1}^{T-1}\sum_{m=1}^{T-t} \mathbb{E}_0 l_{itk_1} l_{itk_2} l_{i,m+t,k_3} \overset{\text{w.p.1}}{\lesssim} \frac{1}{T^3}\sum_{t=1}^{T}\sum_{m=1}^{T} b^m \lesssim \frac{1+d_i}{T^2}.$$

Finally, note that

$$\sum_{t=1}^{T-2}\sum_{s=t+1}^{T-1}\sum_{r=s+1}^{T} |\mathbb{E}_0 l_{itk_1} l_{isk_2} l_{irk_3}|$$

$$= \sum_{t=1}^{T-2}\sum_{s=t+1}^{T-1}\sum_{\Delta_2=1}^{T-s} |\mathbb{E}_0 l_{itk_1} l_{isk_2} l_{i,\Delta_2+s,k_3}| \tag{$\Delta_2 := r-s \geq 1$}$$

$$= \sum_{t=1}^{T-2}\sum_{\Delta_1=1}^{T-1-t}\sum_{\Delta_2=1}^{T-\Delta_1-t} |\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}| \tag{$\Delta_1 := s-t \geq 1$}$$

$$\leq \sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} |\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}|$$

$$= \sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 \leq \Delta_2)|\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}|$$

$$+ \sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 > \Delta_2)|\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}|.$$

Now,

$$\sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 \leq \Delta_2)|\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}|$$

$$= \sum_{\Delta_1=1}^{T}\sum_{q=\Delta_1+1}^{\Delta_1+T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 \leq \Delta_2)|\mathbb{E}_0 l_{i,q-\Delta_1,k_1} l_{iqk_2} l_{i,q+\Delta_2,k_3}| \tag{$q := \Delta_1 + t$}$$

$$\leq \sum_{\Delta_1=1}^{T}\sum_{q=1}^{\Delta_1+T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 \leq \Delta_2)|\mathbb{E}_0 l_{i,q-\Delta_1,k_1} l_{iqk_2} l_{i,q+\Delta_2,k_3}|$$

$$\leq \sum_{q=1}^{2T}\sum_{\Delta_2=1}^{T}\sum_{\Delta_1=1}^{\Delta_2} |\mathbb{E}_0 l_{i,q-\Delta_1,k_1} l_{iqk_2} l_{i,q+\Delta_2,k_3}| \tag{$1 \leq \Delta_1 \leq \Delta_2 \leq T$}$$

$$= \sum_{t=1}^{2T}\sum_{\Delta_2=1}^{T}\sum_{\Delta_1=1}^{\Delta_2} |\mathrm{cov}_0(l_{i,t-\Delta_1,k_1} l_{itk_2}, l_{i,t+\Delta_2,k_3})|.$$

Given $((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$, $l_{i,t-\Delta_1,k_1}$ and l_{itk_2} are $\sigma(Y_{i1},\ldots,Y_{it})$-measurable, whereas $l_{i,t+\Delta_2,k_3}$ is $\sigma(Y_{i,t+\Delta_2},\ldots)$-measurable. Hence, as

$$\mathbb{E}_0(l_{i,t-\Delta_1,k_1} l_{itk_2})^4 \overset{\text{AM-GM}}{\leq} \frac{\mathbb{E}_0 l^8_{i,t-\Delta_1,k_1} + \mathbb{E}_0 l^8_{itk_2}}{2} \overset{\text{(iv)}}{\leq} d_i \quad \& \quad \mathbb{E}_0 l^4_{i,t+\Delta_2,k_3} \overset{\text{(D.1)}}{\leq} 1 + \mathbb{E}_0 l^8_{i,t+\Delta_2,k_3} \overset{\text{(iv)}}{\leq} 1 + d_i,$$

we have that

$$|\mathrm{cov}_0(l_{i,t-\Delta_1,k_1} l_{itk_2}, l_{i,t+\Delta_2,k_3})| \overset{\text{w.p.1}}{\lesssim} (1+d_i)\mathfrak{a}_i^{1/2}(\Delta_2) \tag{Prakasa Rao, Eqn. 63}$$

$$\overset{\text{(iii)}}{\lesssim} (1+d_i)a^{\Delta_2/2}.$$

Consequently, as

$$\sum_{k=1}^{T} kb^k = \frac{b(1-b^T)}{(1-b)^2} - \frac{Tb^{T+1}}{1-b}, \tag{D.3}$$

we have that

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 \le \Delta_2)|\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}| \\
&\qquad \overset{\text{w.p.1}}{\lesssim} (1+d_i)\sum_{t=1}^{2T}\sum_{\Delta_2=1}^{T}\sum_{\Delta_1=1}^{\Delta_2} b^{\Delta_2} && (b := a^{1/2}) \\
&\qquad = (1+d_i)2T\sum_{\Delta_2=1}^{T} \Delta_2 b^{\Delta_2} \\
&\qquad \overset{\text{(D.3)}}{\lesssim} (1+d_i)[\frac{Tb}{(1-b)^2} + \frac{Tb^{T+1}}{(1-b)^2} + \frac{T^2b^{T+1}}{1-b}].
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 > \Delta_2)|\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}| \\
&= \sum_{\Delta_1=1}^{T}\sum_{v=\Delta_1+1}^{\Delta_1+T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 > \Delta_2)|\mathbb{E}_0 l_{i,v-\Delta_2,k_1} l_{i,v+\Delta_1-\Delta_2,k_2} l_{i,v+\Delta_1,k_3}| && (v := \Delta_2 + t) \\
&\le \sum_{\Delta_1=1}^{T}\sum_{v=1}^{\Delta_1+T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 > \Delta_2)|\mathbb{E}_0 l_{i,v-\Delta_2,k_1} l_{i,v+\Delta_1-\Delta_2,k_2} l_{i,v+\Delta_1,k_3}| \\
&\le \sum_{v=1}^{2T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{\Delta_1} |\mathbb{E}_0 l_{i,v-\Delta_2,k_1} l_{i,v+\Delta_1-\Delta_2,k_2} l_{i,v+\Delta_1,k_3}| && (1 \le \Delta_2 < \Delta_1 \le T) \\
&= \sum_{v=1}^{2T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{\Delta_1} |\text{cov}_0(l_{i,v-\Delta_2,k_1}, l_{i,v+\Delta_1-\Delta_2,k_2} l_{i,v+\Delta_1,k_3})|.
\end{aligned}$$

Given $((X_{it})_{t\in\mathbb{N}}, I_{i0}, \alpha_{i0})$, $l_{i,v-\Delta_2,k_1}$ is $\sigma(Y_{i1}, \dots, Y_{i,v-\Delta_2})$-measurable, whereas $l_{i,v+\Delta_1-\Delta_2,k_2}$ and $l_{i,v+\Delta_1,k_3}$ are $\sigma(Y_{i,v+\Delta_1-\Delta_2}, \dots)$-measurable. Hence, as $\mathbb{E}_0(l_{i,v+\Delta_1-\Delta_2,k_2} l_{i,v+\Delta_1,k_3})^4 \overset{\text{AM-GM \& (iv)}}{\lesssim} d_i$ and $\mathbb{E}_0 l^4_{i,v-\Delta_2,k_1} \overset{\text{(D.1) \& (iv)}}{\le} 1 + d_i$, we have

$$\begin{aligned}
&|\text{cov}_0(l_{i,v-\Delta_2,k_1}, l_{i,v+\Delta_1-\Delta_2,k_2} l_{i,v+\Delta_1,k_3})| \overset{\text{w.p.1}}{\lesssim} (1+d_i)\mathfrak{a}_i^{1/2}(\Delta_1) && \text{(Prakasa Rao, Eqn. 63)} \\
&\overset{\text{(iii)}}{\lesssim} (1+d_i)a^{\Delta_1/2}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{T} \mathbb{1}(\Delta_1 > \Delta_2)|\mathbb{E}_0 l_{itk_1} l_{i,\Delta_1+t,k_2} l_{i,\Delta_2+\Delta_1+t,k_3}| \\
&\qquad \overset{\text{w.p.1}}{\lesssim} (1+d_i)\sum_{t=1}^{2T}\sum_{\Delta_1=1}^{T}\sum_{\Delta_2=1}^{\Delta_1} b^{\Delta_1} && (b := a^{1/2}) \\
&\qquad = (1+d_i)2T\sum_{\Delta_1=1}^{T} \Delta_1 b^{\Delta_1} \\
&\qquad \overset{\text{(D.3)}}{\lesssim} (1+d_i)[\frac{Tb}{(1-b)^2} + \frac{Tb^{T+1}}{(1-b)^2} + \frac{T^2b^{T+1}}{1-b}].
\end{aligned}$$

It follows that

$$\frac{1}{T^3}\sum_{t=1}^{T-2}\sum_{s=t+1}^{T-1}\sum_{r=s+1}^{T}|\mathbb{E}_0 l_{itk_1} l_{isk_2} l_{irk_3}| \overset{\text{w.p.1}}{\lesssim} (1+d_i)[\frac{b}{T^2(1-b)^2}+\frac{b^{T+1}}{T^2(1-b)^2}+\frac{b^{T+1}}{T(1-b)}].$$

Combining these results, we obtain that

$$\begin{aligned}
\mathbb{E}_0 l_{\cdot Tk_1} l_{\cdot Tk_2} l_{\cdot Tk_3} &= \frac{1}{n^3}\sum_{i=1}^{n}\mathbb{E}_0 l_{iTk_1} l_{iTk_2} l_{iTk_3} \\
&\overset{\text{w.p.1}}{\lesssim} \frac{1}{n^3}\sum_{i=1}^{n}\frac{(1+d_i)}{T^2}+\frac{1}{n^3}\sum_{i=1}^{n}\frac{(1+d_i)}{T^2}[\frac{b}{T^2(1-b)^2}+\frac{b^{T+1}}{T^2(1-b)^2}+\frac{b^{T+1}}{T(1-b)}] \\
&= \frac{1}{n^2T^2}\frac{1}{n}\sum_{i=1}^{n}(1+d_i)+\frac{1}{n^2T^2}\frac{1}{n}\sum_{i=1}^{n}(1+d_i)[\frac{b}{(1-b)^2}+\frac{b^{T+1}}{(1-b)^2}+\frac{Tb^{T+1}}{(1-b)}] \\
&\overset{\text{(iv)}}{=} O_{\mathrm{p}}(\frac{1}{n^2T^2}), \qquad (\lim_{T\to\infty}Tb^T=0)
\end{aligned}$$

which yields (A.8).

Appendix E. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2022.08.011.

References

Andrews, D.W.K., 1983. First Order Autoregressive Processes and Strong Mixing. Cowles Foundation Discussion Paper No. 664.

Arellano, M., 2003. Discrete choices with panel data. Investig. Econ. XXVII, 423–458.

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

Arellano, M., Hahn, J., 2007. Understanding bias in nonlinear panel models: Some recent developments. In: Blundell, R., Newey, W., Persson, T. (Eds.), Advances in Economics and Econometrics: Ninth World Congress, Vol. 3. Cambridge University Press, Cambridge, UK, pp. 381–409.

Arellano, M., Hahn, J., 2016. A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. Global Econ. Rev. 45, 251–274.

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

Barndorff-Nielsen, O.E., Hall, P., 1988. On the level-error after bartlett adjustment of the likelihood ratio statistic. Biometrika 75, 374–378.

Berger, J.O., Liseo, B., Wolpert, R., 1999. Integrated likelihood functions for eliminating nuisance parameters (with discussion). Statist. Sci. 14, 1–28.

Chamberlain, G., 1980. Analysis of covariance with qualitative data. Rev. Econ. Stud. XLVII, 225–238.

Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 49, 1–39.

Davidson, J., 1994. Stochastic Limit Theory. Oxford University Press, New York, NY, USA.

de Jong, R.M., Woutersen, T., 2011. Dynamic time series binary choice. Econom. Theory 27, 623–702.

Dhaene, G., Sun, Y., 2021. Second-order corrected likelihood for nonlinear panel models with fixed effects. J. Econometrics 220, 227–252.

DiCiccio, T.J., Martin, M.A., Stern, S.E., Young, G.A., 1996. Information bias and adjusted profile likelihoods. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 189–203.

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

Fernández-Val, I., Weidner, M., 2016. Individual and time effects in nonlinear panel models with large N,T. J. Econ. 192, 291–312.

Hahn, J., 1997. A note on the efficient semiparametric estimation of some exponential panel models. Econom. Theory 13, 583–588.

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

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

Halmos, P.R., Savage, L.J., 1949. Applications of the Radon–Nikodym theorem to the theory of sufficient statistics. Ann. Math. Stat. 20, 225–241.

Kalbfleisch, J.D., Sprott, D.A., 1970. Application of likelihood methods to models involving large numbers of parameters (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 32, 175–208.

Kalbfleisch, J.D., Sprott, D.A., 1973. Marginal and conditional likelihoods. Sankhya, Series A 35, 311–328.

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

Li, H., Lindsay, B.G., Waterman, R.P., 2003. Efficiency of projected score methods in rectangular array asymptotics. J. R. Stat. Soc. Ser. B Stat. Methodol. 65, 191–208.

McCullagh, P., Tibshirani, R., 1990. A simple method for the adjustment of profile likelihoods. J. R. Stat. Soc. Ser. B Stat. Methodol. 52, 325–344.

Mykland, P.A., 1999. Bartlett identities and large deviations in likelihood theory. Ann. Statist. 27, 1105–1117.

Neyman, J., Scott, E.L., 1948. Consistent estimation from partially consistent observations. Econometrica 16, 1–32.

Pace, L., Salvan, A., 2006. Adjustments of the profile likelihood from a new perspective. J. Statist. Plann. Inference 136, 3554–3564.

Prakasa Rao, B.L.S., 2009. Conditional independence, conditional mixing and conditional association. Ann. Inst. Stat. Math. 61, 441–460.

Sartori, N., 2003. Modified profile likelihoods in models with stratum nuisance parameters. Biometrika 90, 533–549.

Schumann, M., 2022. Second-order bias reduction for nonlinear panel data models with fixed effects based on expected quantities. Econom. Theory http://dx.doi.org/10.1017/S0266466622000160, (forthcoming).

Schumann, M., Severini, T.A., Tripathi, G., 2021. Integrated likelihood based inference for nonlinear panel data models with unobserved effects. J. Econometrics 223, 73–95.

Severini, T.A., 1998. Likelihood functions for inference in the presence of a nuisance parameter. Biometrika 85, 507–522.

Severini, T.A., 2000. Likelihood Methods in Statistics. Oxford University Press, London, UK.

Severini, T.A., 2007. Integrated likelihood functions for non-Bayesian inference. Biometrika 94, 529–542.

Willink, R., 2003. Relationships between central moments and cumulants, with formulae for the central moments of gamma distributions. Commun. Stat. - Theory Methods 32, 701–704.