Second-order corrected likelihood for nonlinear panel models with fixed effects

Geert Dhaene* Yutao Sun†

January 31, 2018

Abstract

We propose a second-order correction for nonlinear fixed-effect panel models. The correction is made via the log-likelihood function. It removes the two leading terms of the bias of the log-likelihood that arises from estimating the fixed effects. Maximizing the corrected likelihood gives a second-order bias-corrected estimator, with bias $O\left(T^{-3}\right)$, where T is the number of time periods. The corrected likelihood also gives a second-order corrected likelihood ratio statistic. The correction applies to general nonlinear fixed-effect models with independent observations. The bias correction properties are confirmed in simulations for binary-choice models.

JEL Classification: C23

Keywords: Nonlinear panel data model; incidental parameter problem; bias correction; fixed effects

1 Introduction

Panel data are becoming ever more important in economic studies. In panel studies, researchers often attempt to capture individual heterogeneity by introducing individual-specific parameters, or "fixed effects", in the model. When the number of time periods in the data set is small, however, nonlinear models with a large number of fixed-effect parameters may give maximum likelihood (ML) estimates that are severely biased. This is known as the incidental parameter problem (IPP) of Neyman and Scott (1948). Lancaster (2000) surveys the IPP and various approaches trying to solve it. To outline the problem, let the data set have N cross-sectional units, indexed by $i = 1, \cdots, N$, and T time periods, indexed by $t = 1, \cdots, T$. Suppose, as is common in micro-panels, that N is large and T small. This situation is well approximated by asymptotics with $N \to \infty$ and T fixed. Let $\log f\left(Y_{it}; \theta, a_i\right)$ be the log-likelihood associated with observation Y_{it} (possibly conditional on observed covariates X_{it}). Here θ is the vector of parameters that apply to all observations i, t (so θ is the "common" parameter) and $a_1, \cdots, a_N$ are fixed-effect (nuisance) parameters, or "incidental" parameters as Neyman and Scott (1948) called them. Now $\widehat{\alpha}_i$, the ML estimator of a_i, only uses the data from the ith individual. Therefore, when T is fixed, $\widehat{\alpha}_i$ remains random for every i even as

*Faculty of Economics and Business, KU Leuven, Leuven (3000), Belgium. Email: geert.dhaene@kuleuven.be.

†Corresponding author. School of Economics, Northeast Normal University, Changchun (130117), China, and Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam (3000 DR), the Netherlands. Email: sun@ese.eur.nl.

$N \to \infty$. In many models, this introduces a bias in the log-likelihood function in the sense that the ML estimator $\widehat{\theta}$ of θ converges to a point $\theta_T \neq \theta_0$, where θ_0 is the true value of θ. One may view θ_0 as the maximizer of the probability limit, as $N \to \infty$, of the log-likelihood that depends on the true values of the fixed effects; and θ_T as the maximizer of the probability limit, as $N \to \infty$, of the log-likelihood that depends on the estimates of the fixed effects. The two probability limits are in general different, and so are their maximizers, θ_0 and θ_T. The bias of $\widehat{\theta}$ is $O(T^{-1})$. When $T \to \infty$, all $\widehat{\alpha}_i$ converge to their true values and so does $\widehat{\theta}$. However, when $N, T \to \infty$ with T increasing at the same rate as N, the random variation in $\widehat{\alpha}_i$ vanishes too slowly and the limiting distribution of $\widehat{\theta}$ is not centered at θ_0 (see Hahn and Newey 2004).

In the course of nearly seven decades since the IPP was discovered, numerous researchers have attempted to solve it, exactly or approximately, analytically or numerically. In the early years, solutions were specific for a given model or class of models. For example, an exact solution of the IPP is possible if a conditional likelihood exists that depends on θ but not on the fixed effects; see, e.g., Cox (1958) and Andersen (1970). For many models, however, no such conditional likelihood exists. Therefore, researchers have also tried to find general solutions, i.e., solutions that do not depend on the specific functional form of the density. Such solution methods can, in general, only be approximate. They aim at approximate bias correction in some appropriate asymptotic sense. One way of obtaining approximate bias corrections is by model-free methods such as the jackknife or bootstrap. For example, Hahn and Newey (2004) and Dhaene and Jochmans (2015) propose jackknife estimators of the bias. Another way is to analytically derive the approximate bias of $\widehat{\theta}$, as in Hahn and Newey (2004) and Hahn and Kuersteiner (2011), or of the objective function, i.e., the log-likelihood, as in Arellano and Hahn (2016). The bias-corrected log-likelihood of Arellano and Hahn (2016) is an approximation to an infeasible log-likelihood that is not subject to the IPP. The correction reduces the bias in the log-likelihood (normalized by the number of observations) from $O(T^{-1})$ to $O(T^{-2})$. The maximizer of the corrected log-likelihood serves as a bias-corrected estimator, inheriting the order of bias $O(T^{-2})$ from the corrected log-likelihood.

The analytical bias corrections, i.e., the methods based on an explicit formula for the approximate bias, are currently first-order only. They remove one term, the leading term, of a large-T asymptotic expansion of the bias. When T is small, the corrected estimator may still be significantly biased. This is due, partly, to the term of order $O(T^{-2})$ still being non-negligible for small T. This paper derives a refined approximation to the infeasible log-likelihood in order to remove the $O(T^{-2})$ bias term from the log-likelihood as well. The result is a second-order bias-corrected log-likelihood, with bias $O(T^{-3})$. The corrected estimator of θ, obtained from maximizing the corrected log-likelihood, may then serve as a second-order bias-corrected estimator, with bias that is expected to be of order $O(T^{-3})$. Furthermore, it may be conjectured that the asymptotic distribution of the second-order corrected estimator is correctly centered under asymptotics where $N/T^5 \to 0$ (under regularity conditions). As a comparison, the condition for the uncorrected estimator $\widehat{\theta}$ to be asymptotically correctly centered is $N/T \to 0$; and for the first-order corrected estimator the condition is $N/T^3 \to 0$.

We develop the second-order corrected log-likelihood by extending the approach of Arellano and Hahn (2016). The correction can be applied to a general class of models provided that the data are independent and some regularity conditions are satisfied. The corrected log-likelihood depends only on known quantities such as Y_{it}, $\widehat{\alpha}_i$, and log-likelihood derivatives. Hence, it can be constructed in an automated way from the data.

The paper is organized as follows. Section 2 provides a brief review of the IPP and a short introduction on correcting the likelihood function. Section 3 gives the main result, the expression of the second-order corrected log-likelihood, and an outline of the derivation. Section 4 presents examples and simulations. Section 5 concludes. Details of the derivation are given in appendices.

2 Incidental parameter problem and correcting the objective function

We begin with a brief review of the IPP and an introduction on correcting the objective function. In what follows, we assume that expectations exist wherever they are taken and that the stochastic order of remainder terms does not increase on taking expectations. We shall focus on the non-technical aspects here.

2.1 Incidental parameter problem

Let Y_{it} denote the (i,t)th observation, where $i = 1, \cdots, N$ and $t = 1, \cdots, T$. We assume that the Y_{it} are independent across i and t (conditional on covariates and fixed effects). With $f(Y_{it}; \theta, a_i)$ the conditional density of Y_{it}, θ the common parameter, and a_i the fixed-effect parameter, let

$$\alpha_i(\theta) := \arg\max_{a_i} \frac{1}{T}\sum_t \mathbb{E}\log f(Y_{it}; \theta, a_i)$$

and

$$\widehat{\alpha}_i(\theta) := \arg\max_{a_i} \frac{1}{T}\sum_t \log f(Y_{it}; \theta, a_i)$$

be, respectively, the (infeasible) pseudo-true value and the estimator of the fixed effect of the ith individual, for given θ. We use $\mathbb{E}(\cdot)$ to denote the expectation taken under the true density, $f(Y_{it}; \theta_0, \alpha_{i0})$, i.e., the density evaluated at the true values, θ_0 and α_{i0}, of the parameters θ and a_i. Note that $\alpha_i(\theta)$ becomes the true value of the fixed effect when evaluated at $\theta = \theta_0$, i.e., $\alpha_i(\theta_0) = \alpha_{i0}$. Since the purpose is to correct the log-likelihood function at any given θ, we will keep θ fixed in much of the analysis and sometimes drop it as a function argument to condense the notation. We shall also often omit the index i because for each i the log-likelihood is corrected in the same way. Now, for given i and θ, let

$$\widehat{l}(\theta) := \frac{1}{T}\sum_t \log f(Y_{it}; \theta, \widehat{\alpha}_i(\theta))$$

be the profile (or concentrated) log-likelihood function (with i omitted from $\widehat{l}(\theta)$). Throughout, we normalize all log-likelihoods by the number of observations. Averaging $\widehat{l}(\theta)$ over i and maximizing gives the ML estimator, $\widehat{\theta}$. The profile log-likelihood is obtained on plugging $a_i = \widehat{\alpha}_i(\theta)$ into the unprofiled log-likelihood. On the other hand, if $\alpha_i(\theta)$ were known and plugged in, we would obtain the infeasible profile log-likelihood function

$$l(\theta) := \frac{1}{T}\sum_t \log f(Y_{it}; \theta, \alpha_i(\theta)).$$

The infeasible profile log-likelihood is not subject to the IPP. That is,

$$\theta_0 = \operatorname*{plim}_{N\to\infty} \arg\max_{\theta} \frac{1}{N}\sum_i l(\theta)$$

even for fixed T.

The IPP occurs because $\widehat{l}(\theta)$ is a biased estimate of $l(\theta)$. When T is fixed, $\widehat{\alpha}_i(\theta)$ remains random for every i even as $N \to \infty$. In many models, this introduces a bias in the log-likelihood function and the ML estimator of θ, $\widehat{\theta} := \arg\max_\theta N^{-1} \sum_i \widehat{l}(\theta)$, converges to a wrong value, θ_T; i.e., $\text{plim}_{N\to\infty} \widehat{\theta} = \theta_T \neq \theta_0$, where $\theta_T := \text{plim}_{N\to\infty} \arg\max_\theta N^{-1} \sum_i \widehat{l}(\theta)$. When $N, T \to \infty$, the random variation in $\widehat{\alpha}_i(\theta)$ vanishes, so that $\widehat{\theta} \to_p \theta_0$ in general. However, when $N/T \to \kappa$ with $0 < \kappa < \infty$, the limiting distribution of $\sqrt{NT}(\widehat{\theta} - \theta_0)$ is often not centered at 0, implying asymptotic undercoverage of the confidence interval. The bias in the limiting distribution disappears only when $N/T \to 0$, i.e., when T increases faster than N. This result is established in, e.g., Hahn and Kuersteiner (2002) and Hahn and Newey (2004).

The many-normal-means model (Neyman and Scott 1948) is a well-known instance of the IPP. The following example compares $\widehat{l}(\theta)$ and $l(\theta)$, and their maximizers, in this model.

Example 1 (Many normal means) *Assume $Y_{it} \sim \mathcal{N}(\alpha_{i0}, \theta_0)$, where α_{i0}, the mean, may differ across i, and θ_0, the variance, is the same for all i, t. The log-likelihood is, for each i,*

$$\frac{1}{T}\sum\nolimits_t \log f(Y_{it}; \theta, a_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta - \frac{1}{T}\sum_t \frac{(Y_{it} - a_i)^2}{2\theta}.$$

In this example, $\alpha_i(\theta)$ and $\widehat{\alpha}_i(\theta)$ are easily obtained in closed form,

$$\alpha_i := \alpha_i(\theta) = \mathbb{E}Y_{it}, \qquad \widehat{\alpha}_i := \widehat{\alpha}_i(\theta) = \frac{1}{T}\sum_t Y_{it},$$

where α_i and $\widehat{\alpha}_i$ are to be understood as functions of θ. Therefore,

$$\widehat{l}(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta - \frac{1}{T}\sum_t \frac{(Y_{it} - \widehat{\alpha}_i)^2}{2\theta},$$

$$l(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta - \frac{1}{T}\sum_t \frac{(Y_{it} - \alpha_i)^2}{2\theta}.$$

It follows that

$$\widehat{\theta} = \frac{1}{NT}\sum\nolimits_{i,t}(Y_{it} - \widehat{\alpha}_i)^2, \qquad \underset{N\to\infty}{\text{plim}}\, \widehat{\theta} = \theta_0 - \frac{\theta_0}{T}.$$

Hence, $\widehat{\theta}$ is inconsistent for θ_0 when T is fixed, and the bias of $\widehat{\theta}$ is $O(T^{-1})$. On the other hand,

$$\theta_0 = \arg\max_\theta \underset{N\to\infty}{\text{plim}} \frac{1}{N}\sum_i l(\theta) = \underset{N\to\infty}{\text{plim}} \frac{1}{NT}\sum\nolimits_{i,t}(Y_{it} - \alpha_i)^2 = \underset{N\to\infty}{\text{plim}} \frac{1}{NT}\sum\nolimits_{i,t}(Y_{it} - \alpha_{i0})^2,$$

confirming that $l(\theta)$ is not subject to the IPP.

2.2 Correcting the objective function

$l(\theta)$ is infeasible since $\alpha_i(\theta)$ is unknown. However, it is often possible to construct an approximation to $l(\theta)$ using quantities depending only on $\widehat{\alpha}_i(\theta)$, which can be computed from the data. Specifically, if we could build a kth-order approximating function $\widehat{l}^{(k)}(\theta)$, independent of $\alpha_i(\theta)$, that satisfies

$$\mathbb{E}l(\theta) = \mathbb{E}\widehat{l}^{(k)}(\theta) + O(T^{-k-1}),$$

then $\widehat{l}^{(k)}(\theta)$ serves as a corrected log-likelihood function and a less biased estimator of θ_0 may be obtained as

$$\widehat{\theta}^{(k)} := \arg\max_{\theta} \frac{1}{N} \sum_{i} \widehat{l}^{(k)}(\theta). \tag{2.1}$$

Arellano and Hahn (2016) provide the approximating function for $k = 1$, the first order. It takes the form (for a single i)

$$\widehat{l}^{(1)}(\theta) = \widehat{l}(\theta) + \frac{\widehat{b}_1}{T},$$

where $\widehat{b}_1$ is a function evaluated at $\widehat{\alpha}_i(\theta)$. Arellano and Hahn (2016) also show that $\widehat{\theta}^{(1)}$ derived from $\widehat{l}^{(1)}(\theta)$ as in (2.1) is biased to the order $O(T^{-2})$ only. Thus $\widehat{\theta}^{(1)}$ inherits the order of bias reduction from $\widehat{l}^{(1)}(\theta)$. Our contribution is to provide

$$\widehat{l}^{(2)}(\theta) = \widehat{l}(\theta) + \frac{\widehat{b}_1}{T} + \frac{\widehat{b}_2}{T^2}, \tag{2.2}$$

where $\widehat{b}_2$, similar to $\widehat{b}_1$, is a function evaluated at $\widehat{\alpha}_i(\theta)$. Thus, the approximation is refined to the second order.

Three likely consequences of the refinement in (2.2), although not formally investigated here, are that (i) $\widehat{\theta}^{(2)}$ inherits the bias order, $O(T^{-3})$, from $\widehat{l}^{(2)}(\theta)$ (just as $\widehat{\theta}$ and $\widehat{\theta}^{(1)}$ inherit the bias order from $\widehat{l}(\theta)$ and $\widehat{l}^{(1)}(\theta)$); (ii) the asymptotic distribution of $\sqrt{NT}(\widehat{\theta}^{(2)} - \theta_0)$ is normal and centered at 0 under asymptotics where $N/T^5 \to 0$ as $N, T \to \infty$; and (iii) the likelihood ratio statistic based on $\widehat{l}^{(2)}(\theta)$ has an asymptotic χ^2 null distribution under the same asymptotics. Here, (ii) and (iii) are the natural continuation of the conditions $N/T \to 0$ for $\sqrt{NT}(\widehat{\theta} - \theta_0)$ to be normal and centered at 0, and $N/T^3 \to 0$ for $\sqrt{NT}(\widehat{\theta}^{(1)} - \theta_0)$ to be normal and centered at 0. That is, higher-order corrections allow T to grow more slowly with N (in practical terms, T may be small compared to N without causing too much trouble). The focus in this paper is on deriving (2.2). The bias correction property of the corresponding estimator, $\widehat{\theta}^{(2)}$, in particular the conjecture that it inherits the bias order $O(T^{-3})$ from $\widehat{l}^{(2)}(\theta)$, and similarly for the corrected likelihood ratio statistic, will be explored in simulations.

3 Second-order corrected log-likelihood

In this section, we state the main result, the second-order corrected log-likelihood function. We also briefly outline the derivation.

3.1 Main result

We first introduce the following notation.

Notation 1 (Sums of products of individual log-likelihood derivatives) *For given positive integers J and M, let $p_{jm} \in \mathbb{N}$ and $r_{jm} \in \mathbb{N}$ for $j = 1, \cdots, J$ and $m = 1, \cdots, M$. The r_{jm} will indicate orders of derivatives of the log-likelihood w.r.t. a_i. These derivatives will be raised to powers*

p_{jm}, multiplied across j and m, and summed across time. Let

$$R := \begin{pmatrix} r_{11} & \cdots & r_{1M} \\ \vdots & \ddots & \vdots \\ r_{J1} & \cdots & r_{JM} \end{pmatrix}, \qquad P := \begin{pmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & \ddots & \vdots \\ p_{J1} & \cdots & p_{JM} \end{pmatrix},$$
$$\mathcal{T} := \{(t_1, \cdots, t_J) \,|t_j = 1, \cdots, T; t_j \neq t_{j'} \forall j \neq j'; j, j' = 1, \cdots, J\},$$

with the constraints on p_{jm} and r_{jm} that (i) $r_{jm} = 0$ if and only if $p_{jm} = 0$; (ii) if $p_{jm} = 0$ then $p_{jm'} = 0$ for $m' > m$; and (iii) $\sum_{m=1}^M p_{jm} > 0$ and $\sum_{m=1}^M r_{jm} > 0$. Let

$$\mathcal{P}(R, P) := J - \frac{1}{2} \sum\nolimits_{j=1}^J 1 \left(\sum\nolimits_{m=1}^M r_{jm} = 1 \text{ and } \sum\nolimits_{m=1}^M p_{jm} = 1 \right),$$
$$\mathcal{L}(R, P) := \frac{1}{T^{\mathcal{P}(R,P)}} \sum_{(t_1, \cdots, t_J) \in \mathcal{T}} \prod_{j=1}^J \prod_{m=1}^M \left(\nabla_{a_i}^{r_{jm}} \log f\left(Y_{it_j}; \theta, a_i\right)\big|_{a_i = \alpha_i(\theta)} \right)^{p_{jm}},$$

where $1(\cdot)$ is the indicator function. Finally, write

$$\mathcal{L}_{\mathcal{P}}\binom{p_{11}, \cdots, p_{1M}; \cdots; p_{J1}, \cdots p_{JM}}{r_{11}, \cdots, r_{1M}; \cdots; r_{J1}, \cdots r_{JM}} := \mathcal{L}(R, P)$$

to make $\mathcal{P}$ and the p_{jm} and r_{jm} appear explicitly.

Remark 1 (Notation 1) *For every R and P, $\mathcal{P}(R, P)$ is the least half-integer or an integer between $J/2$ and J that ensures that $\mathcal{L}(R, P) = O_p(1)$. We state this as Proposition 3 in Appendix A. Note that $\mathcal{L}(R, P)$ is invariant under permutations of the rows of R and P (the same permutation applied to the rows of R and those of P).*

Example 2 *Let*

$$R = \begin{pmatrix} 1 & 2 \end{pmatrix}, \qquad P = \begin{pmatrix} 2 & 1 \end{pmatrix}.$$

Then $\mathcal{P}(R, P) = 1$ and

$$\mathcal{L}(R, P) = \frac{1}{T} \sum\nolimits_t \left((\nabla_{a_i} \log f(Y_{it}; \theta, a_i))^2 \nabla_{a_i}^2 \log f(Y_{it}; \theta, a_i) \right)_{a_i = \alpha_i(\theta)}.$$

Example 3 *Let*

$$R = \begin{pmatrix} 1 & 0 \\ 2 & 3 \end{pmatrix}, \qquad P = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}.$$

Then

$$\mathcal{L}(R, P) = \frac{1}{T^{\mathcal{P}}} \sum_{t_1 \neq t_2} \left(\nabla_{a_i} \log f(Y_{it_1}; \theta, a_i) \nabla_{a_i}^2 \log f(Y_{it_2}; \theta, a_i) \left(\nabla_{a_i}^3 \log f(Y_{it_2}; \theta, a_i) \right)^2 \right)_{a_i = \alpha_i(\theta)}$$

and

$$\sum\nolimits_{j=1}^J 1 \left(\sum\nolimits_{m=1}^M r_{jm} = 1 \text{ and } \sum\nolimits_{m=1}^M p_{jm} = 1 \right) = 1,$$

so

$$\mathcal{P} = \mathcal{P}(R, P) = \frac{3}{2}.$$

We make the following assumptions about the density in order to justify the definition of $\widehat{l}^{(2)}(\theta)$.

Assumption 1 *For every θ, $\alpha_i(\theta)$ and $\widehat{\alpha}_i(\theta)$ exist. For $r = 1, ..., 4$, $\nabla^r_{a_i} \log f(Y_{it}; \theta, a_i)$ exists and satisfies*

$$\left|\nabla^r_{a_i} \log f(Y_{it}; \theta, a_i)\right| < \infty.$$

Assumption 2 *The second derivative of $\log f(Y_{it}; \theta, a_i)$ w.r.t. a_i satisfies*

$$\frac{1}{T}\sum_t \nabla^2_{a_i} \log f(Y_{it}; \theta, a_i) < 0.$$

Remark 2 (Strict concavity) *In general, strict concavity (Assumption 2) is a strong assumption; see, e.g., Newey and McFadden (1994). However, there are cases where complications arise when it is not assumed. Consider the probit model, and suppose $Y_{it} = 1$ for all t and some i. Then for every θ, $T^{-1}\sum_t \log f(Y_{it}; \theta, a_i)$ is maximized as $a_i \to \infty$ and $\lim_{a_i\to\infty} T^{-1}\sum_t \nabla^2_{a_i} \log f(Y_{it}; \theta, a_i) = 0$. Assumption 2 then amounts to excluding these i (and those i for which $Y_{it} = 0$ for all t). Excluding these i is a natural approach in this model because their profile log-likelihood contribution is identically 0.*

Let

$$l_r := l_r(\theta) := \frac{1}{T}\sum\nolimits_t \nabla^r_{a_i} \log f(Y_{it}; \theta, a_i)\Big|_{a_i=\alpha_i(\theta)},$$
$$\widehat{l}_r := \widehat{l}_r(\theta) := \frac{1}{T}\sum\nolimits_t \nabla^r_{a_i} \log f(Y_{it}; \theta, a_i)\Big|_{a_i=\widehat{\alpha}_i(\theta)}.$$

The main result is as follows.

Proposition 1 (Second-order corrected log-likelihood) *Under Assumptions 1 and 2,*

$$\mathbb{E}l(\theta) = \mathbb{E}\widehat{l}^{(2)}(\theta) + O(T^{-3})$$

where

$$\widehat{l}^{(2)}(\theta) := \widehat{l}(\theta) + \frac{\widehat{b}_1}{T} + \frac{\widehat{b}_2}{T^2} \tag{3.1}$$

with

$$b_1 := \frac{\mathcal{L}_1\binom{2}{1}}{2l_2},$$
$$b_2 := -\frac{\mathcal{L}_2\binom{1,1;1,1}{1,2;1,2}}{l_2^3} - \frac{l_3\mathcal{L}_1\binom{3}{1}}{3l_2^3} - \frac{l_4\mathcal{L}_2\binom{2;2}{1;1}}{12l_2^4} + \frac{5l_4\mathcal{L}_1\binom{2}{1}^2}{24l_2^4} - \frac{5l_3^2\mathcal{L}_1\binom{2}{1}^2}{8l_2^5} + \frac{l_3^2\mathcal{L}_2\binom{2;2}{1;1}}{4l_2^5}$$
$$+ \frac{\mathcal{L}_1\binom{2,1}{1,2}}{l_2^2} - \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{2}{2}}{2l_2^3} - \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{1,1}{1,3}}{2l_2^3} + \frac{3l_3\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{1,1}{1,2}}{2l_2^4},$$

and $\widehat{b}_1$ and $\widehat{b}_2$ are b_1 and b_2 evaluated at $\widehat{\alpha}_i(\theta)$.

Proof. The proof is given in Appendices B.1 and B.2. ∎

The term b_1 and its corresponding estimate $\widehat{b}_1$ coincide with the first-order bias correction term of Arellano and Hahn (2016) for static models. The following example is an application of Proposition 1.

Example 4 (Many normal means, Example 1 continued) *Because* $\nabla^2_{a_i} \log f(Y_{it}; \theta, a_i) = -1/\theta$, *we have*

$$l_2 = -1/\theta, \qquad l_r = 0 \quad \forall r \geq 3, \qquad \mathcal{L}_2\binom{1,1}{1,3} = 0, \qquad \frac{\mathcal{L}_1\binom{2,1}{1,2}}{l_2^2} = \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{2}{2}}{l_2^3} = \frac{\mathcal{L}_1\binom{2}{1}}{l_2},$$

and so

$$\begin{aligned}\widehat{l}^{(2)}(\theta) &= \widehat{l} + \frac{1}{T}\frac{\widehat{\mathcal{L}}_1\binom{2}{1}}{2\widehat{l}_2} - \frac{1}{T^2}\frac{\widehat{\mathcal{L}}_1\binom{1,1;1,1}{1,2;1,2}}{\widehat{l}_2^3} + \frac{1}{T^2}\frac{\widehat{\mathcal{L}}_1\binom{2}{1}}{2\widehat{l}_2} \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta - \frac{1}{2}\left(\frac{1}{T} + \frac{1}{T^2} + \frac{1}{T^3} + \frac{2}{T^4}\right)\sum_t \frac{(Y_{it} - \widehat{\alpha}_i)^2}{\theta},\end{aligned}$$

where $\widehat{l}_2 = -1/\theta$ *and any* $\widehat{\mathcal{L}}$ *is* $\mathcal{L}$ *evaluated at* $\widehat{\alpha}_i$,

$$\begin{aligned}\widehat{\mathcal{L}}_1\binom{2}{1} &= \frac{1}{T}\sum_t \left(\frac{Y_{it} - \widehat{\alpha}_i}{\theta}\right)^2, \\ \widehat{\mathcal{L}}_1\binom{1,1;1,1}{1,2;1,2} &= \frac{1}{T^2}\sum_{t \neq t'}\left(\frac{Y_{it} - \widehat{\alpha}_i}{\theta}\right)\left(\frac{Y_{it'} - \widehat{\alpha}_i}{\theta}\right)\frac{1}{\theta^2} \\ &= -\frac{1}{T^2}\sum_t \frac{(Y_{it} - \widehat{\alpha}_i)^2}{\theta^4}.\end{aligned}$$

It follows that

$$\begin{aligned}\widehat{\theta}^{(2)} &= \left(\frac{1}{T} + \frac{1}{T} + \frac{1}{T^3} + \frac{2}{T^4}\right)\frac{1}{N}\sum_{i,t}(Y_{it} - \widehat{\alpha}_i)^2 \\ &= \left(1 + \frac{1}{T^3} - \frac{2}{T^4}\right)\frac{1}{N(T-1)}\sum_{i,t}(Y_{it} - \widehat{\alpha}_i)^2.\end{aligned}$$

Hence, as $N \to \infty$ *with* T *fixed,*

$$\operatorname*{plim}_{N\to\infty} \widehat{\theta}^{(2)} = \theta_0 + \frac{\theta_0}{T^3} - \frac{2\theta_0}{T^4},$$

so the bias of $\widehat{\theta}^{(2)}$ *is* $O(T^{-3})$.

3.2 Outline of derivation

The derivation of (3.1) follows from stochastic expansions of $\widehat{\alpha}_i(\theta) - \alpha_i(\theta)$ and $\widehat{l}(\theta) - l(\theta)$, similar those in Arellano and Hahn (2016), but with some differences. In Appendix C, we review the derivation of Arellano and Hahn (2016) and highlight the differences.

We first introduce a stochastic expansion of $\widehat{\alpha}_i(\theta) - \alpha_i(\theta)$.

Proposition 2 (Expansion of fixed effect) *Under Assumptions 1 and 2,*

$$\widehat{\alpha}_i(\theta) - \alpha_i(\theta) = -\frac{l_1}{l_2} - \frac{l_1^2 l_3}{2l_2^3} + \frac{l_1^3 l_4}{6l_2^4} - \frac{l_3^2 l_1^3}{2l_2^5} + O_p(T^{-2}). \tag{3.2}$$

Proof. The proof is given in Appendix D. ■

Remark 3 (Bias correction of ML estimator) *Equation (3.2) is not yet accurate enough for the calculation of a second-order corrected ML estimate (viewing $\widehat{\alpha}_i(\theta)$ as an ML estimator) if $\widehat{\alpha}_i(\theta)$ were plugged in on the right-hand side of equation (3.2). This is because of two complications. First, because $\mathbb{E}(\widehat{l}_1/\widehat{l}_2) = \mathbb{E}(l_1/l_2) + O(T^{-1})$, the plug-in version of the right-hand side of (3.2) introduces a bias of order $O(T^{-1})$, which is larger than the targeted $O(T^{-2})$. Second, the $O_p(T^{-2})$ term must also be included to make the second-order corrected ML estimate unbiased to the order $O(T^{-3})$. That is, equation (3.2) has to be extended to an additional order so that the remainder term becomes $O_p(T^{-5/2})$. For the computation of bias-corrected ML estimates, see, e.g., Ferrari, Botter, Cordeiro, and Cribari-Neto (1996). The first complication will be dealt with at a later stage The second one can easily be solved. The technique used to derive equation (3.2) can be continued to produce an arbitrary-order expansion of $\widehat{\alpha}_i(\theta)-\alpha_i(\theta)$. In Appendix E, we give the first 8 terms of the expansion. The representation of the expansion of $\widehat{\alpha}_i(\theta)-\alpha_i(\theta)$ is not unique. Other representations include those of, e.g., Bartlett (1953a,1953b), Haldane and Smith (1956), and Rilstone, Srivastava, and Ullah (1996).*

Consider now the second expansion, that of $\widehat{l}(\theta) - l(\theta)$. Recall that the objects $\widehat{l}(\theta)$ and $l(\theta)$ are identical except that $\widehat{l}(\theta)$ is evaluated at $a_i = \widehat{\alpha}_i(\theta)$ and $l(\theta)$ at $a_i = \alpha_i(\theta)$. Write $\widehat{l} := \widehat{l}(\theta)$ and $l := l(\theta)$ for brevity, given that θ is held fixed. Noting that

$$\widehat{\alpha}_i(\theta) - \alpha_i(\theta) = O_p(T^{-1/2}), \qquad l_1 = O_p(T^{-1/2}), \qquad l_r = O_p(1) \quad \forall r \geq 2,$$

a Taylor expansion of $\widehat{l}$ around $\alpha_i(\theta)$ gives

$$\begin{aligned}\widehat{l} = l &+ l_1(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)) + \frac{1}{2!}l_2(\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^2 + \frac{1}{3!}l_3(\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^3 \\ &+ \frac{1}{4!}l_4(\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^4 + O_p(T^{-5/2}).\end{aligned}$$

Hence, on rearranging,

$$\begin{aligned}l = \widehat{l} &- l_1(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)) - \frac{1}{2}l_2(\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^2 - \frac{1}{6}l_3(\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^3 \\ &- \frac{1}{24}l_4(\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^4 + O_p(T^{-5/2}).\end{aligned} \tag{3.3}$$

The right-hand side of (3.3) depends on the quantity $\widehat{\alpha}_i(\theta) - \alpha_i(\theta)$, for which we use the expansion in Proposition 2. Combining (3.2)–(3.3) and rearranging leads to

$$\mathbb{E}l = \mathbb{E}\widehat{l} + \frac{\mathbb{E}b_1}{T} + \frac{\mathbb{E}b_2'}{T^2} + O(T^{-3}) \tag{3.4}$$

with b_1 as in Proposition 1 whereas $b_2' \neq b_2$; Appendix B.1 gives the derivation and an expression of b_2'. Here b_1 and b_2' are evaluated at $\alpha_i(\theta)$, so that they are not feasible at this stage. In addition, we cannot simply replace $\alpha_i(\theta)$, implicit in b_1, with $\widehat{\alpha}_i(\theta)$ because $\widehat{\alpha}_i(\theta)$ is biased for $\alpha_i(\theta)$, so $\widehat{b}_1$ will be biased for b_1. In particular,

$$\frac{\widehat{b}_1}{T} = \frac{b_1}{T} + O_p(T^{-3/2}),$$

where the remainder term has expectation $O(T^{-2})$, so

$$\mathbb{E}l = \mathbb{E}\widehat{l} + \frac{\mathbb{E}\widehat{b}_1}{T} + \frac{\mathbb{E}b_2'}{T^2} + O(T^{-2}).$$

Here the remainder term is of order $O(T^{-2})$ instead of the targeted $O(T^{-3})$. To deal with this, we apply the same technique as used to derive equation (3.4). Specifically, we Taylor-expand $\widehat{b}_1/T$ around $\alpha_i(\theta)$ to obtain

$$\frac{b_1}{T} = \frac{\widehat{b}_1}{T} - \frac{\nabla b_1}{T}\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right) - \frac{1}{2}\frac{\nabla^2 b_1}{T}\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right)^2 + O_p(T^{-5/2}) \qquad (3.5)$$

where ∇ and ∇^2 denote the first and second derivative w.r.t. $\alpha_i(\theta)$. Next, combining (3.2) and (3.5), we obtain, after rearranging,

$$\frac{b_1}{T} = \frac{\widehat{b}_1}{T} + \frac{b_{1,1}}{T^2} + O_p(T^{-5/2})$$

where $b_{1,1}$ is evaluated at $\alpha_i(\theta)$; see Appendix B.2 for the derivation and an expression of $b_{1,1}$. Then it follows that

$$\mathbb{E}l = \mathbb{E}\widehat{l} + \frac{\mathbb{E}\widehat{b}_1}{T} + \frac{\mathbb{E}b_2}{T^2} + O(T^{-3})$$

where $b_2 = b_2' + b_{1,1}$. Replacing b_2 with $\widehat{b}_2$ introduces a bias of small enough order, $O(T^{-3})$.

4 Simulations

4.1 Bias-corrected estimators

We first present simulations for the fixed-effect logit model. The natural estimator in this model is the conditional logit ML estimator, which is fixed-T consistent (see, e.g., Andersen 1970, Chamberlain 1980, and Heckman 1981). However, a fixed-T consistent estimator in fixed-effect binary-outcome models only exists in the logit model (Chamberlain 2010). Our focus here is on general bias correction methods and we shall only compare the ML estimator, $\widehat{\theta}$, and the first- and second-order corrected estimators, $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$. We generated data from the model

$$Y_{it} = 1\left(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0\right)$$

where ε_{it} is standard-logistically distributed and X_{it} is a scalar covariate. Throughout, we set $N = 10,000$ and let $\theta_0 = \pm 0.5, \pm 1$ and $T = 5, 10, 20$. We used three designs to generate X_{it} and α_i. In Design 1, $X_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i = 0$; in this case the model could be consistently estimated by pooled logit. In Design 2, $X_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \sim \mathcal{N}(0, 1/16)$; here the model could be consistently estimated by random-effect logit with Gaussian effects. In Design 3, $X_{it} \sim \mathcal{N}(\alpha_i, 1)$ with $\alpha_i \sim \mathcal{N}(0, 1/16)$; in this case the model has to be estimated by a fixed-effect method due to the correlation between X_{it} and α_i. Tables 1 to 3 present the simulations results, based on 1,000 Monte Carlo replications. The three designs yield very similar results. The IPP occurs already in Design 1, i.e., it occurs as soon as fixed effects are being estimated, even when their true values are zero. Uniformly across the designs, the ML estimator is heavily biased and the bias is away from

Table 1: Simulations for the logit model, Design 1

Setting	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE
$T = 5$		$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$	
$\widehat{\theta}$	0.6313	26.26%	0.1320	−0.6310	26.21%	0.1318	1.2940	29.40%	0.2946	−1.2928	29.28%	0.2934
$\widehat{\theta}^{(1)}$	0.5383	7.66%	0.0400	−0.5379	7.58%	0.0396	1.0893	8.93%	0.0906	−1.0883	8.83%	0.0895
$\widehat{\theta}^{(2)}$	0.4912	−1.76%	0.0137	−0.4910	−1.79%	0.0138	0.9725	−2.75%	0.0307	−0.9717	−2.83%	0.0309
$T = 10$		$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$	
$\widehat{\theta}$	0.5573	11.46%	0.0579	−0.5570	11.39%	0.0576	1.1229	12.29%	0.1233	−1.1230	12.30%	0.1234
$\widehat{\theta}^{(1)}$	0.5095	1.90%	0.0122	−0.5092	1.84%	0.0120	1.0196	1.96%	0.0216	−1.0197	1.97%	0.0218
$\widehat{\theta}^{(2)}$	0.4993	−0.15%	0.0075	−0.4990	−0.21%	0.0076	0.9965	−0.35%	0.0094	−0.9965	−0.35%	0.0098
$T = 20$		$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$	
$\widehat{\theta}$	0.5267	5.34%	0.0272	−0.5267	5.34%	0.0272	1.0564	5.64%	0.0567	−1.0567	5.67%	0.0571
$\widehat{\theta}^{(1)}$	0.5023	0.45%	0.0056	−0.5022	0.45%	0.0056	1.0044	0.44%	0.0073	−1.0047	0.47%	0.0078
$\widehat{\theta}^{(2)}$	0.4999	−0.02%	0.0050	−0.4999	−0.02%	0.0051	0.9995	−0.05%	0.0058	−0.9998	−0.02%	0.0061

Notes: 1,000 replications. $N = 10,000$. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where ε_{it} is standard-logistically distributed. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate. Design 1: $X_{it} \sim \mathcal{N}(0,1)$, $\alpha_i = 0$.

Table 2: Simulations for the logit model, Design 2

Setting	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE
$T = 5$	$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$		
$\widehat{\theta}$	0.6311	26.21%	0.1318	−0.6307	26.15%	0.1314	1.2941	29.41%	0.2947	−1.2938	29.38%	0.2944
$\widehat{\theta}^{(1)}$	0.5381	7.63%	0.0400	−0.5377	7.53%	0.0393	1.0895	8.95%	0.0907	−1.0892	8.92%	0.0904
$\widehat{\theta}^{(2)}$	0.4907	−1.87%	0.0143	−0.4904	−1.92%	0.0140	0.9721	−2.79%	0.0307	−0.9717	−2.83%	0.0310
$T = 10$	$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$		
$\widehat{\theta}$	0.5574	11.48%	0.0580	−0.5570	11.41%	0.0576	1.1232	12.32%	0.1236	−1.1230	12.30%	0.1235
$\widehat{\theta}^{(1)}$	0.5095	1.91%	0.0120	−0.5092	1.85%	0.0119	1.0198	1.98%	0.0217	−1.0196	1.96%	0.0219
$\widehat{\theta}^{(2)}$	0.4991	−0.18%	0.0072	−0.4988	−0.24%	0.0075	0.9962	−0.38%	0.0095	−0.9960	−0.40%	0.0102
$T = 20$	$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$		
$\widehat{\theta}$	0.5267	5.35%	0.0273	−0.5268	5.36%	0.0274	1.0570	5.70%	0.0574	−1.0568	5.68%	0.0572
$\widehat{\theta}^{(1)}$	0.5023	0.45%	0.0056	−0.5023	0.47%	0.0057	1.0049	0.49%	0.0080	−1.0047	0.47%	0.0078
$\widehat{\theta}^{(2)}$	0.4999	−0.03%	0.0051	−0.4999	−0.01%	0.0052	0.9998	−0.02%	0.0064	−0.9997	−0.03%	0.0062

Notes: 1,000 replications. $N = 10,000$. Model: $Y_{it} = 1\left(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0\right)$ where ε_{it} is standard-logistically distributed. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate. Design 2: $X_{it} \sim \mathcal{N}\left(0,1\right)$ and $\alpha_i \sim \mathcal{N}\left(0,1/16\right)$.

Table 3: Simulations for the logit model, Design 3

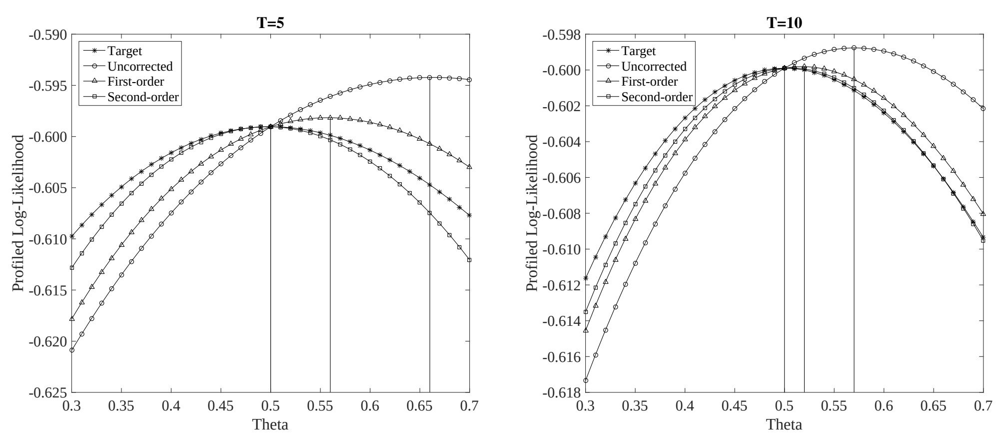
Setting	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE
$T = 5$		$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$	
$\widehat{\theta}$	0.6311	26.22%	0.1318	−0.6309	26.18%	0.1316	1.2946	29.46%	0.2953	−1.2920	29.20%	0.2927
$\widehat{\theta}^{(1)}$	0.5382	7.63%	0.0398	−0.5378	7.57%	0.0395	1.0905	9.05%	0.0916	−1.0884	8.84%	0.0896
$\widehat{\theta}^{(2)}$	0.4904	−1.93%	0.0141	−0.4909	−1.83%	0.0137	0.9709	−2.91%	0.0316	−0.9718	−2.82%	0.0309
$T = 10$		$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$	
$\widehat{\theta}$	0.5573	11.46%	0.0579	−0.5569	11.38%	0.0575	1.1244	12.44%	0.1248	−1.1230	12.30%	0.1234
$\widehat{\theta}^{(1)}$	0.5094	1.88%	0.0120	−0.5091	1.82%	0.0118	1.0205	2.05%	0.0224	−1.0197	1.97%	0.0219
$\widehat{\theta}^{(2)}$	0.4987	−0.26%	0.0074	−0.4988	−0.24%	0.0075	0.9956	−0.44%	0.0099	−0.9965	−0.35%	0.0099
$T = 20$		$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$	
$\widehat{\theta}$	0.5266	5.33%	0.0272	−0.5266	5.33%	0.0272	1.0573	5.73%	0.0577	−1.0567	5.67%	0.0571
$\widehat{\theta}^{(1)}$	0.5021	0.42%	0.0055	−0.5022	0.44%	0.0055	1.0048	0.48%	0.0078	−1.0048	0.48%	0.0077
$\widehat{\theta}^{(2)}$	0.4997	−0.06%	0.0051	−0.4998	−0.04%	0.0050	0.9995	−0.05%	0.0062	−0.9998	−0.02%	0.0060

Notes: $1,000$ replications. $N = 10,000$. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where ε_{it} is standard-logistically distributed. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate. Design 3: $X_{it} \sim \mathcal{N}(\alpha_i, 1)$ and $\alpha_i \sim \mathcal{N}(0, 1/16)$.

zero. When $T = 5$, the bias of $\widehat{\theta}$ varies between 25% and 30%; when $T = 20$, it is still around 5%. The bias-corrected estimators $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ are both very effective in reducing the bias. This holds, in particular, for the second-order bias-corrected estimator, $\widehat{\theta}^{(2)}$. Even when $T = 5$, its bias never exceeds 3%; when $T = 10$, its bias is already uniformly below 0.5%. A glance at the tables also shows that the bias reductions dramatically decrease the root-mean-squared-error (RMSE). The results further show that the expected values of all three estimators are odd functions of θ_0. For example, apart from Monte Carlo error, the reported means of $\widehat{\theta}$ corresponding to $\theta_0 = 0.5$ and $\theta_0 = -0.5$ are equal but with opposite signs. This is consequence of the symmetry of the logit model and of the designs to generate X_{it} and α_i. There seems to be a systematic pattern that the bias of $\widehat{\theta}^{(1)}$ is away from zero and that of $\widehat{\theta}^{(2)}$ is toward zero (note that the tables report percentage bias). We presently do not have an explanation for this phenomenon but, as discussed later, this pattern is not systematic across models and across different values of T. An important point, confirmed by the simulations, is that the bias of $\widehat{\theta}^{(2)}$ is $O(T^{-3})$. This implies, for example, that the bias of $\widehat{\theta}^{(2)}$ should be reduced by a factor roughly equal to 1/8 when T is doubled (and T is not too small).

Figure 1 is a graphical examination of the bias corrections to the log-likelihood in the logit model. It plots the profile log-likelihoods for two data sets, corresponding to $T = 5$ and $T = 10$, both generated with $N = 10,000$, $X_{it} \sim \mathcal{N}(0, 4)$, $\alpha_i = 0$, and $\theta_0 = 0.5$. Letting $X_{it} \sim \mathcal{N}(0, 4)$ introduces enough variation to make the curves sufficiently steep and visually distinguishable from each other. The plotted curves are $N^{-1} \sum_i \widehat{l}(\theta)$ (circles), $N^{-1} \sum_i \widehat{l}^{(1)}(\theta)$ (triangles), $N^{-1} \sum_i \widehat{l}^{(2)}(\theta)$ (squares), and $N^{-1} \sum_i \mathbb{E}l(\theta)$ (asterisks). Here we take $N^{-1} \sum_i \mathbb{E}l(\theta)$ instead of the target log-likelihood, $N^{-1} \sum_i l(\theta)$, for reasons of accuracy. By the convergence of $N^{-1} \sum_i l(\theta)$ to $N^{-1} \sum_i \mathbb{E}l(\theta)$, the two curves should be approximately the same when N is large. However, given the covariate values, $N^{-1} \sum_i \mathbb{E}l(\theta)$ has the advantage of being non-random for finite N. All log-likelihoods are computed for $\theta = 0.3, \cdots, 0.7$ with a step size of 0.01, and the vertical lines indicate the maximizers. A compar-

Figure 1: Profile log-likelihoods for the logit model



Notes: Each figure is computed from a single data set. $N = 10,000$. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where ε_{it} is standard logistic. Data generated with $X_{it} \sim \mathcal{N}(0, 4)$, $\alpha_i = 0$, $\theta_0 = 0.5$. Circles: $N^{-1} \sum_i \widehat{l}(\theta)$; triangles: $N^{-1} \sum_i \widehat{l}^{(1)}(\theta)$; squares: $N^{-1} \sum_i \widehat{l}^{(2)}(\theta)$; asterisks: $N^{-1} \sum_i \mathbb{E}l(\theta)$. All curves are vertically shifted such that they coincide at θ_0. Vertical lines at maximizers.

ison of the two graphs shows that when T increases from 5 to 10, $N^{-1}\sum_i \widehat{l}^{(2)}(\theta)$ converges faster to $N^{-1}\sum_i \mathbb{E}l(\theta)$ than $N^{-1}\sum_i \widehat{l}^{(1)}(\theta)$ and $N^{-1}\sum_i \widehat{l}(\theta)$ do; and that $N^{-1}\sum_i \widehat{l}^{(1)}(\theta)$ converges to $N^{-1}\sum_i \mathbb{E}l(\theta)$ faster than $N^{-1}\sum_i \widehat{l}(\theta)$ does. Even with only $T = 5$ periods, $N^{-1}\sum_i \widehat{l}^{(2)}(\theta)$ is already very accurate, compared to $N^{-1}\sum_i \widehat{l}^{(1)}(\theta)$ and $N^{-1}\sum_i \widehat{l}(\theta)$, as an approximation to $N^{-1}\sum_i \mathbb{E}l(\theta)$ (and to $N^{-1}\sum_i l(\theta)$); and the maximizer of $N^{-1}\sum_i \widehat{l}^{(2)}(\theta)$ is very close to the maximizer of $N^{-1}\sum_i \mathbb{E}l(\theta)$, which is $\theta_0 = 0.5$.

The next example is the fixed-effect probit model; see also Greene, Han, and Schmidt (2002) and Fernández-Val (2009) for simulations in this model. While the IPP in the logit model can be resolved via a conditional likelihood, this is not possible in the probit model. Tables 4 to 6 present simulation results for the probit model under the same designs as for the logit model except that, here, $\varepsilon_{it} \sim \mathcal{N}(0,1)$. The results are qualitatively similar to those for the logit model. Again, the second-order correction is very effective in reducing the bias of the ML estimator. When $T = 5$, the bias of $\widehat{\theta}$ is between 25% and 35% whereas that of $\widehat{\theta}^{(2)}$ is around 4% to 8% (somewhat larger than in the logit model). When $T = 20$, the bias of $\widehat{\theta}$ remains about 5%, while that of $\widehat{\theta}^{(2)}$ is already uniformly below 1% when $T = 10$. There is a symmetry property as θ_0 changes sign, exactly as in the logit model. In the probit model, all three estimators appear to be biased away from zero under the chosen designs, so the opposing signs of the bias of $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$ found in the logit model is not a universal pattern in binary-choice models. One may also notice that, in many cases, the bias of $\widehat{\theta}^{(2)}$ seems to vanish faster than at the rate $O(T^{-3})$. At present, this is not yet fully understood, but note that, for small T, the asymptotics have not fully kicked in yet. In general, the incidental parameter bias of a given estimator may even change sign as T varies. This cannot happen anymore for large enough T because then the leading bias term in the asymptotic expansion determines the sign of the bias (under regularity conditions).

Figure 2 presents plots of the profile log-likelihoods for the probit model, parallelling those for the logit model. Again, $N^{-1}\sum_i \widehat{l}^{(1)}(\theta)$ provides a closer approximation to the target log-likelihood than $N^{-1}\sum_i \widehat{l}(\theta)$ does, and $N^{-1}\sum_i \widehat{l}^{(2)}(\theta)$ provides an even closer appoximation. The main difference, compared with the logit model, is that when $T = 5$, the maximizer of $N^{-1}\sum_i \widehat{l}^{(2)}(\theta)$ is a little less close to $\theta_0 = 0.5$, in line with the somewhat larger bias of $\widehat{\theta}^{(2)}$ found earlier.

Table 4: Simulations for the probit model, Design 1

Setting	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE
$T = 5$	$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$		
$\widehat{\theta}$	0.6239	24.78%	0.1242	−0.6237	24.74%	0.1240	1.3286	32.86%	0.3290	−1.3272	32.72%	0.3275
$\widehat{\theta}^{(1)}$	0.5581	11.62%	0.0587	−0.5579	11.59%	0.0585	1.1846	18.46%	0.1852	−1.1834	18.34%	0.1839
$\widehat{\theta}^{(2)}$	0.5180	3.59%	0.0195	−0.5178	3.56%	0.0193	1.0771	7.71%	0.0781	−1.0761	7.61%	0.0771
$T = 10$	$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$		
$\widehat{\theta}$	0.5504	10.08%	0.0507	−0.5507	10.13%	0.0510	1.1297	12.97%	0.1299	−1.1292	12.92%	0.1295
$\widehat{\theta}^{(1)}$	0.5121	2.42%	0.0130	−0.5124	2.47%	0.0133	1.0378	3.78%	0.0384	−1.0374	3.74%	0.0381
$\widehat{\theta}^{(2)}$	0.5005	0.11%	0.0047	−0.5008	0.16%	0.0049	1.0040	0.40%	0.0077	−1.0036	0.36%	0.0077
$T = 20$	$\theta_0 = 0.5$			$\theta_0 = -0.5$			$\theta_0 = 1$			$\theta_0 = -1$		
$\widehat{\theta}$	0.5229	4.59%	0.0232	−0.5231	4.62%	0.0233	1.0575	5.75%	0.0577	−1.0578	5.78%	0.0580
$\widehat{\theta}^{(1)}$	0.5027	0.53%	0.0043	−0.5028	0.56%	0.0044	1.0084	0.84%	0.0096	−1.0087	0.87%	0.0098
$\widehat{\theta}^{(2)}$	0.4998	−0.04%	0.0034	−0.4999	−0.01%	0.0033	0.9999	−0.01%	0.0046	−1.0002	0.02%	0.0044

Notes: $1,000$ replications. $N = 10,000$. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate. Design 1: $X_{it} \sim \mathcal{N}(0,1)$, $\alpha_i = 0$.

Table 5: Simulations for the probit model, Design 2

Setting	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE
$T=5$		$\theta_0=0.5$			$\theta_0=-0.5$			$\theta_0=1$			$\theta_0=-1$	
$\widehat{\theta}$	0.6266	25.32%	0.1270	−0.6271	25.42%	0.1274	1.3331	33.31%	0.3334	−1.3318	33.18%	0.3321
$\widehat{\theta}^{(1)}$	0.5604	12.08%	0.0609	−0.5608	12.16%	0.0613	1.1888	18.88%	0.1894	−1.1876	18.76%	0.1881
$\widehat{\theta}^{(2)}$	0.5200	4.00%	0.0213	−0.5203	4.07%	0.0216	1.0804	8.04%	0.0813	−1.0793	7.93%	0.0802
$T=10$		$\theta_0=0.5$			$\theta_0=-0.5$			$\theta_0=1$			$\theta_0=-1$	
$\widehat{\theta}$	0.5518	10.36%	0.0521	−0.5518	10.36%	0.0521	1.1311	13.11%	0.1314	−1.1310	13.10%	0.1313
$\widehat{\theta}^{(1)}$	0.5130	2.61%	0.0139	−0.5130	2.61%	0.0140	1.0385	3.85%	0.0392	−1.0385	3.85%	0.0391
$\widehat{\theta}^{(2)}$	0.5012	0.23%	0.0050	−0.5012	0.24%	0.0050	1.0042	0.42%	0.0079	−1.0041	0.41%	0.0079
$T=20$		$\theta_0=0.5$			$\theta_0=-0.5$			$\theta_0=1$			$\theta_0=-1$	
$\widehat{\theta}$	0.5235	4.71%	0.0238	−0.5236	4.71%	0.0238	1.0582	5.82%	0.0585	−1.0585	5.85%	0.0587
$\widehat{\theta}^{(1)}$	0.5029	0.58%	0.0044	−0.5029	0.59%	0.0044	1.0085	0.85%	0.0097	−1.0088	0.88%	0.0100
$\widehat{\theta}^{(2)}$	0.4999	−0.01%	0.0032	−0.5000	−0.01%	0.0033	0.9998	−0.02%	0.0045	−1.0000	0.00%	0.0047

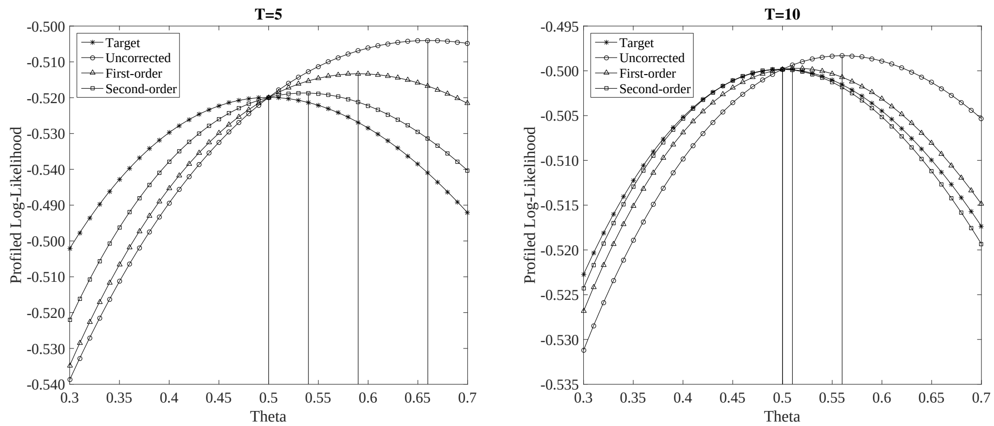
Notes: 1,000 replications. $N = 10,000$. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate. Design 2: $X_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \sim \mathcal{N}(0,1/16)$.

Table 6: Simulations for the probit model, Design 3

Setting	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE	Mean	% Bias	RMSE
$T=5$	$\theta_0=0.5$			$\theta_0=-0.5$			$\theta_0=1$			$\theta_0=-1$		
$\widehat{\theta}$	0.6310	26.20%	0.1313	−0.6241	24.81%	0.1244	1.3445	34.45%	0.3449	−1.3275	32.75%	0.3278
$\widehat{\theta}^{(1)}$	0.5642	12.83%	0.0646	−0.5582	11.64%	0.0588	1.1994	19.94%	0.1999	−1.1837	18.37%	0.1842
$\widehat{\theta}^{(2)}$	0.5233	4.66%	0.0244	−0.5180	3.61%	0.0196	1.0886	8.86%	0.0894	−1.0764	7.64%	0.0773
$T=10$	$\theta_0=0.5$			$\theta_0=-0.5$			$\theta_0=1$			$\theta_0=-1$		
$\widehat{\theta}$	0.5536	10.72%	0.0539	−0.5507	10.15%	0.0510	1.1379	13.79%	0.1382	−1.1290	12.90%	0.1293
$\widehat{\theta}^{(1)}$	0.5142	2.83%	0.0151	−0.5123	2.47%	0.0133	1.0433	4.33%	0.0439	−1.0372	3.72%	0.0379
$\widehat{\theta}^{(2)}$	0.5020	0.40%	0.0054	−0.5007	0.14%	0.0050	1.0073	0.73%	0.0099	−1.0034	0.34%	0.0077
$T=20$	$\theta_0=0.5$			$\theta_0=-0.5$			$\theta_0=1$			$\theta_0=-1$		
$\widehat{\theta}$	0.5243	4.86%	0.0246	−0.5232	4.63%	0.0234	1.0612	6.12%	0.0614	−1.0577	5.77%	0.0579
$\widehat{\theta}^{(1)}$	0.5033	0.65%	0.0047	−0.5028	0.56%	0.0042	1.0098	0.98%	0.0109	−1.0086	0.86%	0.0097
$\widehat{\theta}^{(2)}$	0.5001	0.03%	0.0033	−0.4999	−0.02%	0.0032	1.0004	0.04%	0.0046	−1.0001	0.01%	0.0044

Notes: $1,000$ replications. $N = 10,000$. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate. Design 3: $X_{it} \sim \mathcal{N}(\alpha_i, 1)$ and $\alpha_i \sim \mathcal{N}(0, 1/16)$.

Figure 2: Profile log-likelihoods for the probit model



Notes: Each figure is computed from a single data set. $N = 10,000$. Model: $Y_{it} = 1\left(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0\right)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Data generated with $X_{it} \sim \mathcal{N}(0,4)$, $\alpha_i = 0$, $\theta_0 = 0.5$. Circles: $N^{-1}\sum_i \widehat{l}(\theta)$; triangles: $N^{-1}\sum_i \widehat{l}^{(1)}(\theta)$; squares: $N^{-1}\sum_i \widehat{l}^{(2)}(\theta)$; asterisks: $N^{-1}\sum_i \mathbb{E}l(\theta)$. All curves are vertically shifted such that they coincide at θ_0. Vertical lines at maximizers.

Our last example is the Poisson model, where Y_{it} is Poisson distributed with mean $\exp\left(X_{it}\theta_0 + \alpha_i\right)$. In this model, there is no IPP (see Lancaster 2002 and Blundell, Griffith, and Windmeijer 2002), so it is of interest to examine the effect of (unnecessary) bias correction. In fact, the bias correction is expected to introduce bias, although only of order $O(T^{-2})$ and $O(T^{-3})$ for $\widehat{\theta}^{(1)}$ and $\widehat{\theta}^{(2)}$, respectively. We generated data $Y_{it} \sim \text{Poisson}(\exp\left(X_{it}\theta_0 + \alpha_i\right))$ with $X_{it} \sim \mathcal{U}(0,1)$, $\alpha_i = 0$, $\theta_0 = 0.5$, $N = 10,000$, and $T = 10, 20$. Table 7 presents the results, based on 1,000 Monte Carlo replications. The simulations confirm the theoretical predictions. The bias correction introduces a small bias and leaves the RMSE virtually unchanged.

Table 7: Simulations for the Poisson model

	Mean	% Bias	RMSE	Mean	% Bias	RMSE
		$T = 10$			$T = 20$	
$\widehat{\theta}$	0.4997	−0.05%	0.0099	0.5002	0.04%	0.0069
$\widehat{\theta}^{(1)}$	0.4984	−0.33%	0.0100	0.4998	−0.04%	0.0069
$\widehat{\theta}^{(2)}$	0.4969	−0.61%	0.0103	0.4996	−0.09%	0.0069

Notes: 1,000 replications. $N = 10,000$. Model: $Y_{it} \sim\text{Poisson}(\exp\left(X_{it}\theta_0 + \alpha_i\right))$. Data generated with $X_{it} \sim \mathcal{U}(0,1)$, $\alpha_i = 0$, $\theta_0 = 0.5$. $\widehat{\theta}$ is the ML estimate, $\widehat{\theta}^{(1)}$ is the first-order corrected estimate, $\widehat{\theta}^{(2)}$ is the second-order corrected estimate.

4.2 Bias-corrected LR statistics

Log-likelihood corrections may also be used for improving inference based on the likelihood ratio (LR) test. We compare four versions of the LR statistic: the standard LR statistic, which is based on $\widehat{l}(\theta)$; the first- and second-order corrected LR statistics, based on $\widehat{l}^{(1)}(\theta)$ and $\widehat{l}^{(2)}(\theta)$; and the infeasible LR statistic, based on $l(\theta)$. For fixed T, only the infeasible LR statistic is asymptotically χ^2 distributed, as $N \to \infty$, under the null. All three feasible LR tests are distorted under the null, with rejection probabilities converging to 1 as $N \to \infty$. For finite N, one expects the LR test based on $\widehat{l}(\theta)$ to be the most heavily distorted and that based on $\widehat{l}^{(2)}(\theta)$ to be the least distorted.

We ran simulations for the LR statistics in the context of the fixed-effect probit model under Design 3 (i.e., with $X_{it} \sim \mathcal{N}(\alpha_i, 1)$ and $\alpha_i \sim \mathcal{N}(0, 1/16)$) and with $\theta_0 = 0.5$. Tables 8 and 9 present the rejection probabilities of the LR test when the nominal level is 5%, for null hypotheses ranging from $\theta_0 = 0.4$ to $\theta_0 = 0.6$, the true value being $\theta_0 = 0.5$ throughout. The rejection probabilities are expressed as percentages and are based on 10,000 Monte Carlo replications. Table 8 gives the results for $N = 100$ and Table 9 for $N = 1,000$. All theoretical predictions are confirmed by the simulations. The standard LR test heavily overrejects: when $N = 100$ and $T = 5$, the probality of rejecting $H_0 : \theta_0 = 0.5$ is 40%; when $T = 20$, it is still 12%. For the first-order corrected LR test, these probabilities improve to 20% ($T = 5$) and 6% ($T = 20$), and for the second-order corrected test they further improve to 11% ($T = 5$) and 6% ($T = 20$). When N is increased to $1,000$, the distortions get worse for the standard LR test, with rejection probabilities of 100% ($T = 5$) and 61% ($T = 20$). They also get worse, when $N = 1,000$, for the bias-corrected LR tests, especially when $T = 5$, with rejection probabilities of 81% (first-order correction) and 25% (second-order correction), but much less so or not at all when $T = 20$, with rejection probabilities of 7% (first-order correction) and 5% (second-order correction). It is also of interest to consider the power of the various tests, i.e., the rejection probabilities corresponding to incorrect null hypotheses, at least for tests with only mild distortions when the null holds. Here we see, in particular, that for $T \geq 10$ the second-order corrected LR test gives rejection probabilities that are close to those of the infeasible LR statistic, uniformly over the tested values of θ_0. This is also reflected very clearly in the Figures 3 and 4, which are visual presentations of the Tables 8 and 9 for $T = 5, 10$. The overall conclusion of the simulations for the probit model in this subsection is that the LR statistic greatly benefits from bias correction of the profile log-likelihood, and that the second-order correction improves on the first-order correction.

Table 8: Rejection probability (in %) of LR test at 5% level (probit model, $N = 100$)

H_0	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.54	0.55	0.56	0.57	0.58	0.59	0.6
											$T = 5$										
$\widehat{l}(\theta)$	83	80	76	72	68	63	59	54	49	45	40	36	31	28	24	22	19	16	14	12	11
$\widehat{l}^{(1)}(\theta)$	67	62	57	51	46	41	36	31	27	24	20	18	15	13	11	11	10	10	11	12	13
$\widehat{l}^{(2)}(\theta)$	50	44	39	33	28	25	21	18	15	13	11	10	9	10	10	12	13	16	19	22	26
$l(\theta)$	43	37	30	24	20	15	12	9	7	6	5	5	6	7	9	12	15	19	24	29	35
											$T = 10$										
$\widehat{l}(\theta)$	86	81	76	69	62	54	47	39	32	25	20	16	12	10	8	7	7	8	10	13	16
$\widehat{l}^{(1)}(\theta)$	70	62	54	45	37	30	23	18	14	11	8	7	7	8	10	14	18	23	29	36	44
$\widehat{l}^{(2)}(\theta)$	62	53	45	36	29	22	17	13	10	8	7	7	8	11	15	19	25	31	38	47	55
$l(\theta)$	65	56	47	37	28	21	15	11	8	6	5	6	7	10	14	19	26	34	42	51	59
											$T = 20$										
$\widehat{l}(\theta)$	96	92	87	79	70	59	47	36	26	18	12	8	6	6	8	12	19	27	36	47	58
$\widehat{l}^{(1)}(\theta)$	89	82	73	62	50	38	27	18	12	8	6	6	8	14	21	30	41	52	63	73	81
$\widehat{l}^{(2)}(\theta)$	88	80	70	59	46	35	24	16	10	7	5	6	10	16	24	33	45	56	67	77	84
$l(\theta)$	90	82	73	61	48	35	24	16	10	6	5	6	9	15	24	34	46	58	69	79	87

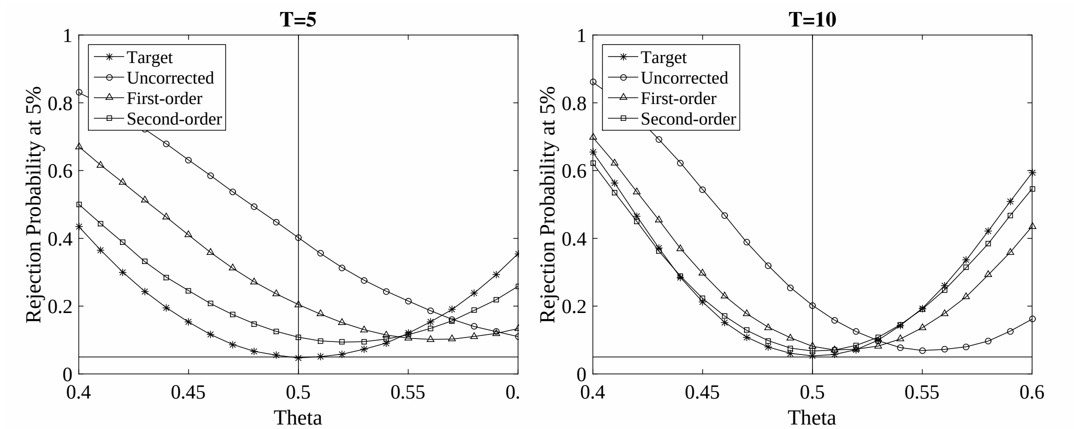
Notes: 10,000 replications. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Data generated with $X_{it} \sim \mathcal{N}(\alpha_i, 1)$, $\alpha_i = \mathcal{N}(0, 1/16)$, $\theta_0 = 0.5$, $N = 100$. The entries are rejection probabilities, in %, of the LR test of H_0 given in the first row, calculated based on the log-likelihood given in the first column.

Table 9: Rejection probability (in %) of LR test at 5% level (probit model, $N = 1000$)

H_0	0.40	0.41	0.42	0.43	0.44	0.45	0.46	0.47	0.48	0.49	0.5	0.51	0.52	0.53	0.54	0.55	0.56	0.57	0.58	0.59	0.6
											$T = 5$										
$\widehat{l}(\theta)$	100	100	100	100	100	100	100	100	100	100	100	100	99	97	93	87	78	66	52	39	27
$\widehat{l}^{(1)}(\theta)$	100	100	100	100	100	100	99	98	96	90	81	68	53	38	24	15	10	11	16	26	40
$\widehat{l}^{(2)}(\theta)$	100	100	100	99	98	93	85	73	57	40	25	14	9	10	18	30	45	61	76	86	93
$l(\theta)$	100	100	100	99	95	85	68	47	27	12	5	5	12	26	46	67	83	93	98	99	100
											$T = 10$										
$\widehat{l}(\theta)$	100	100	100	100	100	100	100	100	99	97	90	76	54	32	15	7	8	19	39	61	79
$\widehat{l}^{(1)}(\theta)$	100	100	100	100	100	99	95	84	62	37	18	7	8	21	43	66	85	95	99	100	100
$\widehat{l}^{(2)}(\theta)$	100	100	100	100	99	94	82	58	33	15	6	10	25	49	73	89	97	99	100	100	100
$l(\theta)$	100	100	100	100	99	96	84	60	32	12	5	10	28	56	80	94	99	100	100	100	100
											$T = 20$										
$\widehat{l}(\theta)$	100	100	100	100	100	100	100	100	98	88	61	26	7	8	30	64	90	98	100	100	100
$\widehat{l}^{(1)}(\theta)$	100	100	100	100	100	100	99	90	61	25	7	10	36	72	93	99	100	100	100	100	100
$\widehat{l}^{(2)}(\theta)$	100	100	100	100	100	100	97	83	50	17	5	16	48	81	97	100	100	100	100	100	100
$l(\theta)$	100	100	100	100	100	100	98	85	52	17	5	17	50	84	98	100	100	100	100	100	100

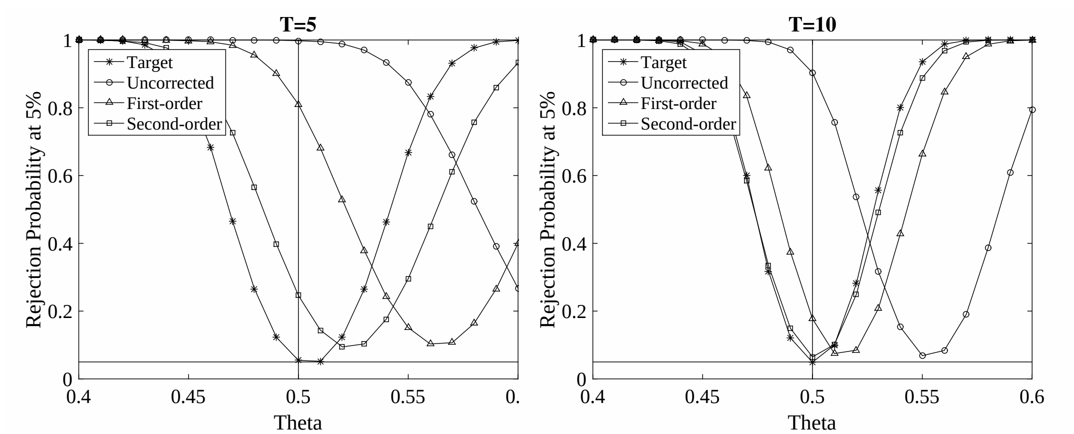
Notes: 10,000 replications. Model: $Y_{it} = 1(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Data generated with $X_{it} \sim \mathcal{N}(\alpha_i, 1)$, $\alpha_i = \mathcal{N}(0, 1/16)$, $\theta_0 = 0.5$, $N = 1,000$. The entries are rejection probabilities, in %, of the LR test of H_0 given in the first row, calculated based on the log-likelihood given in the first column.

Figure 3: Rejection probability (in %) of LR test at 5% level (probit model, $N = 100$)



Notes: 10,000 replications. Model: $Y_{it} = 1\left(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0\right)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Data generated with $X_{it} \sim \mathcal{N}(\alpha_i, 1)$, $\alpha_i = \mathcal{N}(0, 1/16)$, $\theta_0 = 0.5$, $N = 100$. H_0 on the horizontal axis. Vertical line at $\theta_0 = 0.5$ and horizontal line at 5%. LR tests based on $\widehat{l}(\theta)$ (circles), $\widetilde{l}^{(1)}(\theta)$ (triangles), $\widetilde{l}^{(2)}(\theta)$ (squares), and $l(\theta)$ (asterisks).

Figure 4: Rejection probability (in %) of LR test at 5% level (probit model, $N = 1000$)



Notes: 10,000 replications. Model: $Y_{it} = 1\left(X_{it}\theta_0 + \alpha_i + \varepsilon_{it} \geq 0\right)$ where $\varepsilon_{it} \sim \mathcal{N}(0,1)$. Data generated with $X_{it} \sim \mathcal{N}(\alpha_i, 1)$, $\alpha_i = \mathcal{N}(0, 1/16)$, $\theta_0 = 0.5$, $N = 1,000$. H_0 on the horizontal axis. Vertical line at $\theta_0 = 0.5$ and horizontal line at 5%. LR tests based on $\widehat{l}(\theta)$ (circles), $\widetilde{l}^{(1)}(\theta)$ (triangles), $\widetilde{l}^{(2)}(\theta)$ (squares), and $l(\theta)$ (asterisks).

5 Concluding remarks

We derived a second-order bias correction for the profile log-likelihood in nonlinear fixed-effect panel models. The correction removes the first two terms of an expansion of the bias that arises from estimating the fixed effects. As a result, the bias of the profile log-likelihood, normalized by the number of observations, is reduced from $O(T^{-1})$ to $O(T^{-3})$. Simulations in binary-choice models show that the corresponding maximizer of the corrected log-likelihood inherits the order of bias reduction.

The bias correction was developed under independence of the observations (conditional on covariates and fixed effects). When the observations are dependent, the IPP is typically much more severe. This is already manifest in the linear model, where least-squares is not subject to the IPP under independence, while is it seriously biased in the AR(1) model (Nickell 1981). Therefore, it would be of great interest to generalize the second-order correction to the case of dependent data, that is, to extend Arellano and Hahn (2016) to the second order.

Given that the second-order bias correction substantially improves on the first-order correction even for small T, one may wonder what the third-order and, possibly, arbitrary-order corrections might attain in small-T samples. It should, in principle, be possible to extend the bias-correction approach to any order.

Appendices

A Preliminaries

Let R_j and P_j be the j-th row of R and P, and let

$$s_{it}\left(R_j, P_j\right) := \prod\nolimits_{m=1}^{M} \left(\nabla_{a_i}^{r_{jm}} \log f\left(Y_{it}; \theta, a_i\right)\big|_{a_i=\alpha_i(\theta)} \right)^{p_{jm}},$$
$$\mathcal{S}\left(R, P\right) := \sum_{(t_1,\cdots,t_J)\in\mathcal{T}} \prod_{j=1}^{J} s_{it_j}\left(R_j, P_j\right).$$

Proposition 3 (Stochastic order of $\mathcal{L}\left(R, P\right)$) *Let*

$$\mathcal{J}_0\left(R, P\right) := \left\{ j | 1 \leq j \leq J \text{ and } \sum\nolimits_{m=1}^{M} r_{jm} = 1 \text{ and } \sum\nolimits_{m=1}^{M} p_{jm} = 1 \right\},$$
$$\mathcal{J}_1\left(R, P\right) := \left\{1, \cdots, J\right\} \backslash \mathcal{J}_0\left(R, P\right).$$

Suppose

$$\prod_{j\in\mathcal{J}_1(R,P)} \prod_{m=1}^{M} \left(\nabla_{a_i}^{r_{jm}} \log f\left(Y_{it_j}; \theta, a_i\right)\big|_{a_i=\alpha_i(\theta)} \right)^{p_{jm}} \tag{5.1}$$

is nonconstant (i.e., stochastic) for every $r_{jm} \leq 4$ and

$$\mathbb{E}\left(\prod_{j\in\mathcal{J}_1(R,P)} \prod_{m=1}^{M} \left(\nabla_{a_i}^{r_{jm}} \log f\left(Y_{it_j}; \theta, a\right)\big|_{a_i=\alpha_i(\theta)} \right)^{p_{jm}} \right) \neq 0. \tag{5.2}$$

Then $\mathcal{P}\left(R, P\right)$ is the least half-integer or integer such that $\mathcal{L}\left(R, P\right) = O_p\left(1\right)$.

Proof. As $T \to \infty$, the random variable $s_{it}\left(R_j, P_j\right)$ satisfies

$$\frac{1}{T} \sum\nolimits_{t} s_{it}\left(R_j, P_j\right) \xrightarrow{p} \mathbb{E}s_{it}\left(R_j, P_j\right),$$

$$\frac{1}{T}\sum\nolimits_t s_{it}\left(R_j,P_j\right)=\mathbb{E}s_{it}\left(R_j,P_j\right)+O_p(T^{-1/2}).$$

Therefore, when $j\in\mathcal{J}_1\left(R,P\right)$, we have $\mathbb{E}s_{it}\left(R_j,P_j\right)\neq 0$ and

$$\sum\nolimits_t s_{it}\left(R_j,P_j\right)=O_p\left(T\right);$$

whereas, when $j\in\mathcal{J}_0\left(R,P\right)$, we have $\mathbb{E}s_{it}\left(R_j,P_j\right)=0$ and

$$\sum\nolimits_t s_{it}\left(R_j,P_j\right)=O_p(T^{1/2}).$$

Now

$$\mathcal{S}\left(R,P\right)=\sum_{(t_1,\cdots,t_J)\in\mathcal{T}}\left(\prod_{j\in\mathcal{J}_0(R,P)}s_{it_j}\left(R_j,P_j\right)\right)\left(\prod_{j\in\mathcal{J}_1(R,P)}s_{it_j}\left(R_j,P_j\right)\right)$$

is a J-fold summation with each fold being $\sum_t s_{it}\left(R_j,P_j\right)$. Therefore,

$$\begin{aligned}\mathcal{S}\left(R,P\right)&=\left(\prod_{j\in\mathcal{J}_0(R,P)}O_p(T^{1/2})\right)\left(\prod_{j\in\mathcal{J}_1(R,P)}O_p\left(T\right)\right)\\&=O_p\left(T^{\frac{|\mathcal{J}_0(R,P)|}{2}}\right)O_p\left(T^{|\mathcal{J}_1(R,P)|}\right)\\&=O_p\left(T^{|\mathcal{J}_1(R,P)|+\frac{1}{2}|\mathcal{J}_0(R,P)|}\right).\end{aligned}$$

Since $\mathcal{P}\left(R,P\right)=J-\frac{1}{2}\left|\mathcal{J}_0\left(R,P\right)\right|=\left|\mathcal{J}_1\left(R,P\right)\right|+\frac{1}{2}\left|\mathcal{J}_0\left(R,P\right)\right|$, it is obvious that

$$\mathcal{L}\left(R,P\right)=\frac{1}{T^{J-\frac{1}{2}|\mathcal{J}_0(R,P)|}}\mathcal{S}\left(R,P\right)=O_p\left(1\right).$$

■

Remark 4 (Stochastic order) *When (5.1) is constant or when (5.2) is not satisfied, $\mathcal{L}\left(R,P\right)$ is still $O_p\left(1\right)$, but $\mathcal{P}\left(R,P\right)$ is no longer the least half-integer or integer such that $\mathcal{L}\left(R,P\right)=O_p\left(1\right)$.*

Lemma 1 (Expectation of $\mathcal{L}\left(R,P\right)$) *$\mathbb{E}\mathcal{L}\left(R,P\right)=0$ if $\mathcal{P}\left(R,P\right)<J$. In addition, when (5.1) non-constant and (5.2) holds, $\mathbb{E}\mathcal{L}\left(R,P\right)=0$ if and only if $\mathcal{P}\left(R,P\right)<J$.*

Proof. $\mathcal{P}\left(R,P\right)<J$ is equivalent to $\left|\mathcal{J}_0\left(R,P\right)\right|>0$. Hence

$$\mathcal{S}\left(R,P\right)=\sum_{(t_1,\cdots,t_J)\in\mathcal{T}}\left(\prod_{j\in\mathcal{J}_1(R,P)}s_{it_j}\left(R_j,P_j\right)\right)\left(\prod_{j\in\mathcal{J}_0(R,P)}s_{it_j}\left(R_j,P_j\right)\right).$$

By independence,

$$\mathbb{E}\mathcal{S}\left(R,P\right)=\sum_{(t_1,\cdots,t_J)\in\mathcal{T}}\left(\prod_{j\in\mathcal{J}_1(R,P)}\mathbb{E}s_{it_j}\left(R_j,P_j\right)\right)\left(\prod_{j\in\mathcal{J}_0(R,P)}\mathbb{E}s_{it_j}\left(R_j,P_j\right)\right),$$

where, since $\mathbb{E}s_{it}\left(R_j,P_j\right)=0$ if and only if $j\in\mathcal{J}_0\left(R,P\right)$, $\mathbb{E}\mathcal{S}\left(R,P\right)=0$ and $\mathbb{E}\mathcal{L}\left(R,P\right)=0$ if and only if $\mathcal{P}\left(R,P\right)<J$. ■

Proposition 4 (Product of $\mathcal{L}(R,P)$) *Let*

$$R := \begin{pmatrix} r_{11} & \cdots & r_{1M} \\ \vdots & \ddots & \vdots \\ r_{J1} & \cdots & r_{JM} \end{pmatrix}, \qquad P := \begin{pmatrix} p_{11} & \cdots & p_{1M} \\ \vdots & \ddots & \vdots \\ p_{J1} & \cdots & p_{JM} \end{pmatrix},$$
$$R' := \begin{pmatrix} r'_{11} & \cdots & r'_{1M'} \\ \vdots & \ddots & \vdots \\ r'_{J'1} & \cdots & r'_{J'M'} \end{pmatrix}, \qquad P' := \begin{pmatrix} p'_{11} & \cdots & p'_{1M'} \\ \vdots & \ddots & \vdots \\ p'_{J'1} & \cdots & p'_{J'M'} \end{pmatrix},$$

where, without loss of generality, $J' \leq J$. Let $\binom{A}{B}$ and (A,B) denote, respectively, the vertical and horizontal stacks of two matrices A and B of conformable dimensions. For $j = 1, \cdots, J$ and $j' = 1, \cdots, J'$, let

$$c_j := \binom{P_j}{R_j}, \qquad c'_{j'} := \binom{P'_{j'}}{R'_{j'}}, \qquad \langle c_j, c'_{j'} \rangle := \binom{P_j, P'_{j'}}{R_j, R'_{j'}},$$

in which any pairs (p_{jm}, r_{jm}) and $(p'_{j'm'}, r'_{j'm'})$ are removed if, respectively, $p_{jm} = r_{jm} = 0$ and $p'_{j'm'} = r'_{j'm'} = 0$. For every integer $z = 0, \cdots, J'$ and given positive integers $j_1, \cdots, j_z \leq J$ and $j'_1, \cdots, j'_z \leq J'$ with $j_u \neq j_v$ for all $u \neq v$ and $j'_u \neq j'_v$ for all $u \neq v$, let

$$\begin{aligned} \mathcal{C}_{j \neq j_1, \cdots, j_z} &:= \{c_j | j = 1, \cdots, J; j \neq j_1, \cdots, j_z\}, \\ \mathcal{C}'_{j \neq j'_1, \cdots, j'_z} &:= \left\{c'_j | j = 1, \cdots, J'; j \neq j'_1, \cdots, j'_z\right\}. \end{aligned}$$

Let

$$\mathcal{S}(c_1, \cdots, c_J) := \mathcal{S}(R, P), \qquad \mathcal{S}(c'_1, \cdots, c'_{J'}) := \mathcal{S}(R', P').$$

Then, using the notation defined in Remark 5 below,

$$\begin{aligned} &\mathcal{S}(c_1, \cdots, c_J)\mathcal{S}(c'_1, \cdots, c'_{J'}) \\ &= \sum_{z \in (0, \cdots, J')} \sum_{\substack{j_1 < \cdots < j_z \in (1, \cdots, J) \\ j'_1, \cdots, j'_z \in (1, \cdots, J')}} \mathcal{S}\left(\langle c_{j_1}, c'_{j'_1} \rangle, \cdots, \langle c_{j_z}, c'_{j'_z} \rangle, \mathcal{C}_{j \neq j_1, \cdots, j_z}, \mathcal{C}'_{j \neq j'_1, \cdots, j'_z}\right), \end{aligned}$$

where

$$\mathcal{S}\left(\langle c_{j_1}, c'_{j'_1} \rangle, \cdots, \langle c_{j_z}, c'_{j'_z} \rangle, \mathcal{C}_{j \neq j_1, \cdots, j_z}, \mathcal{C}'_{j \neq j'_1, \cdots, j'_z}\right) = 0$$

if

$$\begin{aligned} &T < z + |\mathcal{C}_{j \neq j_1, \cdots, j_z}| + \left|\mathcal{C}'_{j \neq j'_1, \cdots, j'_z}\right| = J + J' - z, \\ &z < J + J' - T. \end{aligned}$$

Proof. It follows from the definitions that

$$\begin{aligned} \mathcal{S}(c_1, \cdots, c_J)\mathcal{S}(c'_1, \cdots, c'_{J'}) & \\ = &\mathcal{S}(c_1, \cdots, c_J, c'_1, \cdots, c'_{J'}) \\ &+ \mathcal{S}(\langle c_1, c'_1 \rangle, \cdots, c_J, c'_2, \cdots, c'_{J'}) + \cdots + \mathcal{S}(c_1, \cdots, \langle c_J, c'_1 \rangle, c'_2, \cdots, c'_{J'}) \\ &+ \cdots \end{aligned}$$

$$+\mathcal{S}\left(\langle c_1, c'_{J'}\rangle, \cdots, c_J, c'_1, \cdots, c'_{J'-1}\right) + \cdots + \mathcal{S}\left(c_1, \cdots, \langle c_J, c'_{J'}\rangle, c'_1, \cdots, c'_{J'-1}\right)$$
$$+ \sum_{\substack{j_1<j_2\in(1,\cdots,J)\\ j'_1,j'_2\in\left(1,\cdots,J'\right)}} \mathcal{S}\left(c_1, \cdots, \langle c_{j_1}, c'_{j'_1}\rangle, \cdots, \langle c_{j_2}, c'_{j'_2}\rangle, \cdots, c_J, \mathcal{C}'_{j\neq j'_1,j'_2}\right)$$
$$+ \sum_{\substack{j_1<j_2<j_3\in(1,\cdots,J)\\ j'_1,j'_2,j'_3\in\left(1,\cdots,J'\right)}} \mathcal{S}\left(c_1, \cdots, \langle c_{j_1}, c'_{j'_1}\rangle, \cdots, \langle c_{j_3}, c'_{j'_3}\rangle, \cdots, c_J, \mathcal{C}'_{j\neq j'_1,j'_2,j'_3}\right)$$
$$+ \sum_{z\in(4,\cdots,J')} \sum_{\substack{j_1<\cdots<j_z\in(1,\cdots,J)\\ j'_1,\cdots,j'_z\in\left(1,\cdots,J'\right)}} \mathcal{S}\left(\langle c_{j_1}, c'_{j'_1}\rangle, \cdots, \langle c_{j_z}, c'_{j'_z}\rangle, \mathcal{C}_{j\neq j_1,\cdots,j_z}, \mathcal{C}'_{j\neq j'_1,\cdots,j'_z}\right),$$

which leads to Proposition 4. ■

Remark 5 ($\mathcal{C}$ and $\mathcal{C}'$) *$\mathcal{C}_{j\neq j_1,\cdots,j_z}$ and $\mathcal{C}'_{j\neq j'_1,\cdots,j'_z}$ serve as collections of c_j and $c'_{j'}$ for $j \neq j_1, \cdots, j_z$ and $j' \neq j'_1, \cdots, j'_z$. Note that $j_1, \cdots, j_z$ need not be identical to $j'_1, \cdots, j'_z$ and, even when they are identical, $\mathcal{C}_{j\neq j_1,\cdots,j_z}$ and $\mathcal{C}'_{j\neq j'_1,\cdots,j'_z}$ may still be different. Also, for every nonnegative integer $n \leq J'$, we use the notation*

$$\mathcal{S}(c_1, \cdots, c_J) := \mathcal{S}(c_1, \cdots, c_n, \mathcal{C}_{j\neq 1,\cdots,n}),$$
$$\mathcal{S}(c'_1, \cdots, c'_{J'}) := \mathcal{S}(c'_1, \cdots, c'_n, \mathcal{C}'_{j\neq 1,\cdots,n}).$$

Remark 6 (Fixed T calculation) *For any T,*

$$\mathcal{S}(c_1, \cdots, c_J)\mathcal{S}(c'_1, \cdots, c'_{J'}) = \sum_{\max(J+J'-T,0)\leq z\leq J'} \sum_{\substack{j_1<\cdots<j_z\in(1,\cdots,J)\\ j'_1,\cdots,j'_z\in\left(1,\cdots,J'\right)}} \mathcal{S}\left(\begin{array}{c}\langle c_{j_1}, c'_{j'_1}\rangle, \cdots, \langle c_{j_z}, c'_{j'_z}\rangle,\\ \mathcal{C}_{j\neq j_1,\cdots,j_z}, \mathcal{C}'_{j\neq j'_1,\cdots,j'_z}\end{array}\right).$$

B Derivation of corrected log-likelihood (proof of Proposition 1)

The corrected log-likelihood is derived in Appendices B.1 and B.2, following the outline given in Section 3.2. Appendix B.3 gives details on an algebraic procedure used in the derivation. Appendices B.4 and B.5 contain some further intermediate steps of the calculations in Appendices B.1 and B.2, respectively.

B.1 Main derivation, part 1

We first present a way to derive the expansion $\mathbb{E}l = \mathbb{E}\widehat{l} + \mathbb{E}b_1/T + \mathbb{E}b'_2/T^2 + O(T^{-3})$, where $\widehat{l} := \widehat{l}\,(\theta)$, $l := l\,(\theta)$, $b_1 = O_p\,(1)$, and $b'_2 = O_p\,(1)$. Here b_1 and b_2 are evaluated at $\alpha_i\,(\theta)$, so the expansion is not yet accurate enough.

From Proposition 2 and equation (3.3), it follows that

$$l = \widehat{l} \underbrace{+\frac{l_1^2}{2l_2}}_{[A]} \underbrace{+\frac{l_3 l_1^3}{6l_2^3}}_{[B]} \underbrace{+\frac{l_3^2 l_1^4}{8l_2^5}}_{[C]} \underbrace{-\frac{l_4 l_1^4}{24 l_2^4}}_{[D]} + O_p(T^{-5/2}). \tag{5.3}$$

Next, we expand each ratio in (5.3) as a power series in $1/T$ and drop the terms that have zero expectation in the resulting expansion. We perform this expansion in two steps. First, we expand the products of sums in the numerators of $[A]$ to $[D]$ as a series of additive terms. This is a rearrangement similar to $(a+b)(c+d) = ac+ad+bc+bd$. We describe this procedure in Appendix B.3. It gives

$$[A] = \frac{\mathcal{L}_1\binom{2}{1}}{2Tl_2} + \underbrace{\frac{\mathcal{L}_1\binom{1;1}{1;1}}{2Tl_2}}_{[E]},$$

$$[B] = \underbrace{\frac{\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{2T^{1.5}l_2^3}}_{[F]} + \underbrace{\frac{\mathcal{L}_{2.5}\binom{1;1;1;1}{1;1;1;3}}{6T^{1.5}l_2^3}}_{[G]} + \frac{\mathcal{L}_2\binom{2;1,1}{1;1,3}}{2T^2l_2^3} + \underbrace{\frac{\mathcal{L}_2\binom{1;1;1,1}{1;1;1,3}}{2T^2l_2^3}}_{[H]} + \frac{l_3\mathcal{L}_1\binom{3}{1}}{6T^2l_2^3} + O_p(T^{-5/2}),$$

$$[C] = \underbrace{\frac{3\mathcal{L}_4\binom{1;1;2;1;1}{1;1;1;3;3}}{4T^2l_2^5}}_{[I]} + \underbrace{\frac{\mathcal{L}_4\binom{1;1;1;1;1;1}{1;1;1;1;3;3}}{8T^2l_2^5}}_{[J]} + \frac{l_3^2\mathcal{L}_1\binom{2}{1}^2}{8T^2l_2^5} + \frac{l_3^2\mathcal{L}_2\binom{2;2}{1;1}}{4T^2l_2^5} + O_p(T^{-5/2}),$$

$$[D] = \underbrace{-\frac{\mathcal{L}_3\binom{1;1;2;1}{1;1;1;4}}{4T^2l_2^4}}_{[K]} \underbrace{-\frac{\mathcal{L}_3\binom{1;1;1;1;1}{1;1;1;1;4}}{24T^2l_2^4}}_{[L]} - \frac{l_4\mathcal{L}_2\binom{2;2}{1;1}}{12T^2l_2^4} - \frac{l_4\mathcal{L}_1\binom{2}{1}^2}{24T^2l_2^4} + O_p(T^{-5/2}).$$

Second, we drop the ratios that have zero expectation to the order that matters. Here it can be verified, by Lemma 1, that $[H]$ to $[L]$ have zero expectation to the order $o(T^{-2})$, so that they can be dropped. For example,

$$\begin{aligned}\mathbb{E}\frac{\mathcal{L}_2\binom{1;1;1,1}{1;1;1,3}}{2T^2l_2^3} &= \frac{1}{T^2}\frac{\mathbb{E}\mathcal{L}_2\binom{1;1;1,1}{1;1;1,3}}{2\mathbb{E}\left(l_2^3\right)} + o(T^{-2}) \\ &= o(T^{-2}),\end{aligned}$$

hence the term $[H]$ can be dropped. On the other hand, $[E]$, $[F]$, and $[G]$ need to be investigated further. This is because, while the leading terms in their stochastic expansions have zero expectation, the next terms may have nonzero expectation of order $O(T^{-2})$, so they need to be included. In particular,

$$[E] = E' + O_p(T^{-3/2}), \qquad [F] = F' + O_p(T^{-2}), \qquad [G] = G' + O_p(T^{-2}),$$

where $\mathbb{E}E' = \mathbb{E}F' = \mathbb{E}G' = 0$. Here, terms that are $O_p(T^{-3/2})$ or $O_p(T^{-2})$, which are of lower order than $O_p(T^{-5/2})$, must be included in the derivation if they have nonzero expectation, whereas the terms E', F', and G' can be dropped. To find these terms, first observe that, as $l_2 = \mathbb{E}l_2 + O_p(T^{-1/2})$, $1/l_2$ can be expanded, giving

$$\frac{1}{l_2} = \frac{1}{\mathbb{E}l_2} - \frac{1}{(\mathbb{E}l_2)^2}\left(l_2 - \mathbb{E}l_2\right) + \frac{1}{(\mathbb{E}l_2)^3}\left(l_2 - \mathbb{E}l_2\right)^2 + O_p(T^{-3/2}). \tag{5.4}$$

Remark 7 (Expansion (5.4)) *The properties of the expansion in (5.4) are studied by, e.g., Rice (2008). It is known that the Taylor series of the reciprocal function is only convergent in a specific region, which, in our setting, is* $2\mathbb{E}l_2 < l_2 < 0$. *This, however, does not prevent using (5.4), since* $l_2 \to_p \mathbb{E}l_2$ *as* $T \to \infty$.

Next, we replace $1/l_2$ in $[E]$, $[F]$, and $[G]$ by the right-hand side of (5.4). Using the procedure of Appendix B.3, this results in

$$\begin{aligned}
[E] &= \frac{\mathcal{L}_1\binom{1;1}{1;1}}{2T}\left(\frac{1}{\mathbb{E}l_2} - \frac{1}{(\mathbb{E}l_2)^2}(l_2-\mathbb{E}l_2) + \frac{1}{(\mathbb{E}l_2)^3}(l_2-\mathbb{E}l_2)^2 + O_p(T^{-3/2})\right)\\
&= \frac{3\mathcal{L}_1\binom{1;1}{1;1}}{2T\mathbb{E}l_2} - \frac{3\mathcal{L}_2\binom{1;1;1}{1;1;2}}{2T(\mathbb{E}l_2)^2} + \frac{\mathcal{L}_3\binom{1;1;1;1}{1;1;2;2}}{2T(\mathbb{E}l_2)^3} - \frac{3\mathcal{L}_{1.5}\binom{1;1,1}{1;1,2}}{T^{1.5}(\mathbb{E}l_2)^2} + \frac{2\mathcal{L}_{2.5}\binom{1;1;1,1}{1;2;1,2}}{T^{1.5}(\mathbb{E}l_2)^3}\\
&\quad \underbrace{+\frac{\mathcal{L}_2\binom{1,1;1,1}{1,2;1,2}}{T^2(\mathbb{E}l_2)^3}}_{[E.1]} + \frac{\mathcal{L}_2\binom{1;1;2}{1;1;2}}{2T^2(\mathbb{E}l_2)^3} + O_p(T^{-5/2}),
\end{aligned}$$

where only $[E.1]$ has nonzero expectation, apart from the $O_p(T^{-5/2})$ remainder term. Similarly,

$$\begin{aligned}
[F] &= \frac{\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{2T^{1.5}}\left(\frac{1}{\mathbb{E}l_2} - \frac{1}{(\mathbb{E}l_2)^2}(l_2-\mathbb{E}l_2) + \frac{1}{(\mathbb{E}l_2)^3}(l_2-\mathbb{E}l_2)^2 + O_p(T^{-3/2})\right)^3\\
&= \frac{2\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{T^{1.5}(\mathbb{E}l_2)^3} - \frac{3\mathcal{L}_{3.5}\binom{1;2;1;1}{1;1;3;2}}{2T^{1.5}(\mathbb{E}l_2)^4} \underbrace{- \frac{3\mathcal{L}_3\binom{2;1;1,1}{1;3;1,2}}{2T^2(\mathbb{E}l_2)^4}}_{[F.1]} + O_p(T^{-5/2}),\\
[G] &= \frac{\mathcal{L}_{2.5}\binom{1;1;1;1}{1;1;1;3}}{6T^{1.5}}\left(\frac{1}{\mathbb{E}l_2} - \frac{1}{(\mathbb{E}l_2)^2}(l_2-\mathbb{E}l_2) + \frac{1}{(\mathbb{E}l_2)^3}(l_2-\mathbb{E}l_2)^2 + O_p(T^{-3/2})\right)^3\\
&= \frac{2\mathcal{L}_{2.5}\binom{1;1;1;1}{1;1;1;3}}{3T^{1.5}(\mathbb{E}l_2)^3} - \frac{\mathcal{L}_{3.5}\binom{1;1;1;1;1}{1;1;1;3;2}}{2T^{1.5}(\mathbb{E}l_2)^4} - \frac{3\mathcal{L}_3\binom{1;1;1;1,1}{1;1;3;1,2}}{2T^2(\mathbb{E}l_2)^4} + O_p(T^{-5/2}),
\end{aligned}$$

where only $[F.1]$ has nonzero expectation, apart from the $O_p(T^{-5/2})$ remainders. Now, we further drop $[G]$ and replace $[E]$ and $[F]$ with $[E.1]$ and $[F.1]$, so that

$$\mathbb{E}l = \mathbb{E}\widehat{l} + \frac{\mathbb{E}b_1}{T} + \frac{\mathbb{E}b_2'}{T^2} + O(T^{-3}) \tag{5.5}$$

where

$$\begin{aligned}
b_1 &:= \frac{\mathcal{L}_1\binom{2}{1}}{2l_2},\\
b_2' &:= \frac{\mathcal{L}_2\binom{1,1;1,1}{1,2;1,2}}{(\mathbb{E}l_2)^3} - \frac{3\mathcal{L}_3\binom{2;1;1,1}{1;3;1,2}}{2(\mathbb{E}l_2)^4} + \frac{\mathcal{L}_2\binom{2;1,1}{1;1,3}}{2l_2^3} + \frac{l_3\mathcal{L}_1\binom{3}{1}}{6l_2^3}\\
&\quad - \frac{l_4\mathcal{L}_2\binom{2;2}{1;1}}{12l_2^4} - \frac{l_4\mathcal{L}_1\binom{2}{1}^2}{24l_2^4} + \frac{l_3^2\mathcal{L}_1\binom{2}{1}^2}{8l_2^5} + \frac{l_3^2\mathcal{L}_2\binom{2;2}{1;1}}{4l_2^5}.
\end{aligned}$$

Here it can be verified that $b_1 = O_p(1)$ and $b_2' = O_p(1)$.

B.2 Main derivation, part 2

Next, we need to account for the fact that $\widehat{b}_1$ is biased for b_1. Taylor-expanding $\widehat{b}_1/T$ around $\alpha_i(\theta)$ gives

$$\begin{aligned}\frac{\widehat{b}_1}{T} &= \frac{b_1}{T} + \frac{1}{T}\nabla b_1\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right) + \frac{1}{2}\frac{1}{T}\nabla^2 b_1\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right)^2 + O_p(T^{-5/2}),\\ \frac{b_1}{T} &= \frac{\widehat{b}_1}{T} - \frac{1}{T}\nabla b_1\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right) - \frac{1}{2}\frac{1}{T}\nabla^2 b_1\left(\widehat{\alpha}_i(\theta) - \alpha_i(\theta)\right)^2 + O_p(T^{-5/2}),\end{aligned}$$

where ∇^r denotes the r-th derivative w.r.t. $\alpha_i(\theta)$. Using Proposition 2 and rearranging,

$$\begin{aligned}\frac{b_1}{T} = \frac{\widehat{b}_1}{T} &+ \frac{l_1\mathcal{L}_1\binom{1,1}{1,2}}{Tl_2^2} - \frac{l_1^2\mathcal{L}_1\binom{2}{2}}{2Tl_2^3} - \frac{l_1^2\mathcal{L}_1\binom{1,1}{1,3}}{2Tl_2^3} + \frac{l_1^2 l_4\mathcal{L}_1\binom{2}{1}}{4Tl_2^4}\\ &+ \frac{3l_1^2 l_3\mathcal{L}_1\binom{1,1}{1,2}}{2Tl_2^4} - \frac{3l_1^2 l_3^2\mathcal{L}_1\binom{2}{1}}{4Tl_2^5} - \frac{l_1 l_3\mathcal{L}_1\binom{2}{1}}{2Tl_2^3} + O_p(T^{-5/2}).\end{aligned}$$

Now we apply the procedure described in Appendix B.1 to the object b_1/T instead of l; see Appendix B.5 for the intermediate steps. The result is

$$\frac{\mathbb{E}b_1}{T} = \frac{\mathbb{E}\widehat{b}_1}{T} + \frac{\mathbb{E}b_{1,1}}{T^2} + O(T^{-3}) \tag{5.6}$$

with

$$\begin{aligned}b_{1,1} := \frac{3\mathcal{L}_3\binom{2;1;1,1}{1;3;1,2}}{2\left(\mathbb{E}l_2\right)^4} - \frac{2\mathcal{L}_2\binom{1,1;1,1}{1,2;1,2}}{\left(\mathbb{E}l_2\right)^3} + \frac{\mathcal{L}_1\binom{2,1}{1,2}}{l_2^2} - \frac{\mathcal{L}_2\binom{2;1,1}{1;1,3}}{2l_2^3} - \frac{3l_3^2\mathcal{L}_1\binom{2}{1}^2}{4l_2^5}\\ - \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{2}{2}}{2l_2^3} - \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{1,1}{1,3}}{2l_2^3} - \frac{l_3\mathcal{L}_1\binom{3}{1}}{2l_2^3} + \frac{l_4\mathcal{L}_1\binom{2}{1}^2}{4l_2^4} + \frac{3l_3\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{1,1}{1,2}}{2l_2^4}.\end{aligned}$$

Combining (5.5) and (5.6) gives Proposition 1.

B.3 Expansion of a product of sums

Here we introduce the algorithm used to expand the products of sums involved in the derivation. Given a positive integer U and given $\mathcal{L}(R^{(u)}, P^{(u)})$, $u = 1, \cdots, U$, we find positive integers V and W_v, a series of $\mathcal{L}(R^{'(v,w)}, P^{'(v,w)})$, $v = 1, \cdots, V$ and $w = 1, \cdots, W_v$, and positive integers $p^{(v)}$, $v = 1, \cdots, V$, such that

$$\prod_{u=1}^{U}\mathcal{L}(R^{(u)}, P^{(u)}) = \sum_{v=1}^{V}\left(\frac{1}{T^{p^{(v)}}}\prod_{w=1}^{W_v}\mathcal{L}(R^{'(v,w)}, P^{'(v,w)})\right) \tag{5.7}$$

and, for every v, exactly one of the following conditions is satisfied.

Condition 1 $\mathbb{E}\mathcal{L}(R^{'(v,w)}, P^{'(v,w)}) \neq 0$ *for every* $w = 1, \cdots, W_v$.

Condition 2 $\mathbb{E}\prod_{w=1}^{W_v}\mathcal{L}(R^{'(v,w)}, P^{'(v,w)}) = 0$ *only if* $W_v = 1$.

The numbers $p^{(v)}$ are determined by the restriction that $\mathcal{L}(R,P)$ be $O_p(1)$. Note that $\mathcal{L}(R^{(u)},P^{(u)})$ and $\mathcal{L}(R^{(u')},P^{(u')})$ can be identical when $u \neq u'$. Similarly, $\mathcal{L}(R'^{(v,w)},P'^{(v,w)})$ and $\mathcal{L}(R'^{\left(v',w'\right)},P'^{\left(v',w'\right)})$ can be identical when $v \neq v'$ or $w \neq w'$. Note, further, that every l_r can be rewritten as $\mathcal{L}(R^{(u)},P^{(u)})$, so products of several l_r can also be accommodated. We first illustrate the expansion in a simple case.

Example 5 *It is easy to see that*

$$\begin{aligned} l_1^2 &= \frac{1}{T^2}\sum_t \left(\nabla_{a_i}\log f\left(Y_{it};\theta,a_i\right)\right)^2\Big|_{a_i=\alpha_i(\theta)} \\ &\quad + \frac{1}{T^2}\sum_{t_1\neq t_2}\nabla_{a_i}\log f\left(Y_{it_1};\theta,a_i\right)\nabla_{a_i}\log f\left(Y_{it_2};\theta,a_i\right)\big|_{a_i=\alpha_i(\theta)} \\ &= \frac{1}{T}\mathcal{L}_1\tbinom{2}{1} + \frac{1}{T}\mathcal{L}_1\tbinom{1;1}{1;1}. \end{aligned}$$

Here $p^{(v)}=1$ *and* $W_v=1$ *for* $v=1,2$.

The expansion is carried out iteratively. First, we take any two $\mathcal{L}(R^{(u)},P^{(u)})$ that satisfy $\mathbb{E}\mathcal{L}(R^{(u)},P^{(u)})=0$ and calculate the product. The condition $\mathbb{E}\mathcal{L}(R^{(u)},P^{(u)})=0$ can be checked using Lemma 1 in Appendix A. The calculation of the product of two such $\mathcal{L}(R^{(u)},P^{(u)})$ is given in Proposition 4 in Appendix A. If there is only one $\mathcal{L}(R^{(u)},P^{(u)})$ satisfying $\mathbb{E}\mathcal{L}(R^{(u)},P^{(u)})=0$, we calculate the product of this particular $\mathcal{L}(R^{(u)},P^{(u)})$ and any other $\mathcal{L}(R^{(u')},P^{(u')})$. The following example illustrates the first step of the iteration.

Example 6 *We compute* $l_3 l_1^3$ *as*

$$l_3 l_1^3 = \frac{l_3 l_1 \mathcal{L}_1\tbinom{2}{1}}{T} + \frac{l_3 l_1 \mathcal{L}_1\tbinom{1;1}{1;1}}{T}.$$

Here $\mathbb{E}l_1=0$*, so the factors* l_1 *and* l_1 *(i.e.,* l_1^2*) have to be processed whereas the other factor,* $l_3 l_1$*, is kept unchanged.*

In every step of the iteration, the expression from the preceding step is taken as input and the procedure is repeated. We stop if the resulting expression satisfies exactly one of the Conditions 1 and 2. The following example illustrates the continuation of the algorithm.

Example 7 *Continuing the above example, the next step computes*

$$\begin{aligned} l_1\mathcal{L}_1\tbinom{2}{1} &= \frac{1}{T}\mathcal{L}_1\tbinom{3}{1} + \frac{1}{T^{1/2}}\mathcal{L}_{1.5}\tbinom{2;1}{1;1}, \\ l_1\mathcal{L}_1\tbinom{1;1}{1;1} &= \frac{1}{T^{1/2}}\mathcal{L}_{1.5}\tbinom{1;1;1}{1;1;1} + \frac{2}{T^{1/2}}\mathcal{L}_{1.5}\tbinom{2;1}{1;1}. \end{aligned}$$

Hence

$$l_3 l_1^3 = \frac{l_3\mathcal{L}_1\tbinom{3}{1}}{T^2} + \frac{3 l_3\mathcal{L}_{1.5}\tbinom{2;1}{1;1}}{T^{1.5}} + \frac{l_3\mathcal{L}_{1.5}\tbinom{1;1;1}{1;1;1}}{T^{1.5}}.$$

Here, the first term on the right-hand side satisfies Condition 1 and, hence, needs no further processing. The two other terms need to be expanded further.

Remark 8 (Uniqueness of expansion and check of the algorithm) *The final expression for the expansion may depend on the order (i.e., the selection) of the factors that are processed in the intermediate calculations. However, all variants are equivalent in that, for any model and data set, when evaluated at any given θ, they give the same value. One may also stop the algorithm when $W_v = 1$ for all v, regardless of Conditions 1 and 2. This yields an equivalent expansion for the given value of T which can be of interest when T is small. When T is large, however, this way of calculating the expansion delivers too many terms, slowing down the computation. We implemented the expansion procedure as a symbolic computer algorithm and verified symbolically that the algorithm works as desired. In addition, we also checked that the left-hand side of equation (5.7), the input of the algorithm, is numerically identical to the right-hand side, the output, for various U and $\mathcal{L}(R^{(u)}, P^{(u)})$.*

B.4 Intermediate steps in step 1

The calculation regarding $[A]$ is straightforward. For $[B]$ to $[D]$, it can be derived that

$$
\begin{aligned}
[B] &= \frac{\mathcal{L}_1\binom{2}{1} l_3 \mathcal{L}_{0.5}\binom{1}{1}}{6T^{1.5} l_2^3} + \frac{l_3 \mathcal{L}_{0.5}\binom{1}{1}\mathcal{L}_1\binom{1;1}{1;1}}{6T^{1.5} l_2^3} = \frac{l_3 \mathcal{L}_1\binom{3}{1}}{6T^2 l_2^3} + \frac{l_3 \mathcal{L}_{1.5}\binom{1;2}{1;1}}{2T^{1.5} l_2^3} + \frac{l_3 \mathcal{L}_{1.5}\binom{1;1;1}{1;1;1}}{6T^{1.5} l_2^3} \\
&= \frac{\mathcal{L}_2\binom{2;1,1}{1;1,3}}{2T^2 l_2^3} + \frac{\mathcal{L}_2\binom{1;1;1,1}{1;1;1,3}}{2T^2 l_2^3} + \frac{\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{2T^{1.5} l_2^3} + \frac{\mathcal{L}_{2.5}\binom{1;1;1;1}{1;1;1;3}}{6T^{1.5} l_2^3} + \frac{l_3 \mathcal{L}_1\binom{3}{1}}{6T^2 l_2^3} + O_p(T^{-5/2}),
\end{aligned}
$$

$$
\begin{aligned}
[C] &= \frac{\mathcal{L}_1\binom{2}{1} l_3^2 \mathcal{L}_{0.5}\binom{1}{1}^2}{8T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_{0.5}\binom{1}{1}^2 \mathcal{L}_1\binom{1;1}{1;1}}{8T^2 l_2^5} \\
&= \frac{\mathcal{L}_1\binom{2}{1}^2 l_3^2}{8T^2 l_2^5} + \frac{\mathcal{L}_1\binom{2}{1} l_3^2 \mathcal{L}_1\binom{1;1}{1;1}}{8T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_{0.5}\binom{1}{1}\mathcal{L}_{1.5}\binom{2;1}{1;1}}{4T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_{0.5}\binom{1}{1}\mathcal{L}_{1.5}\binom{1;1;1}{1;1;1}}{8T^2 l_2^5} \\
&= \frac{\mathcal{L}_1\binom{2}{1}^2 l_3^2}{8T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_2\binom{2;2}{1;1}}{4T^2 l_2^5} + \frac{3 l_3^2 \mathcal{L}_2\binom{1;1;2}{1;1;1}}{4T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_2\binom{1;1;1;1}{1;1;1;1}}{8T^2 l_2^5} + O_p(T^{-5/2}) \\
&= \frac{\mathcal{L}_1\binom{2}{1}^2 l_3^2}{8T^2 l_2^5} + \frac{3 l_3 \mathcal{L}_3\binom{1;1;2;1}{1;1;1;3}}{4T^2 l_2^5} + \frac{l_3 \mathcal{L}_3\binom{1;1;1;1;1}{1;1;1;1;3}}{4T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_2\binom{2;2}{1;1}}{2T^2 l_2^5} \\
&\quad + \frac{\mathcal{L}_1\binom{2}{1}^2 l_3^2}{8T^2 l_2^5} + \frac{3 l_3 \mathcal{L}_3\binom{1;1;2;1}{1;1;1;3}}{4T^2 l_2^5} + O_p(T^{-5/2}) \\
&= \frac{3\mathcal{L}_4\binom{1;1;2;1;1}{1;1;1;3;3}}{4T^2 l_2^5} + \frac{\mathcal{L}_4\binom{1;1;1;1;1;1}{1;1;1;1;3;3}}{8T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_1\binom{2}{1}^2}{8T^2 l_2^5} + \frac{l_3^2 \mathcal{L}_2\binom{2;2}{1;1}}{4T^2 l_2^5} + O_p(T^{-5/2}),
\end{aligned}
$$

and

$$
\begin{aligned}
[D] &= -\frac{\mathcal{L}_1\binom{2}{1} l_4 \mathcal{L}_{0.5}\binom{1}{1}^2}{24T^2 l_2^4} - \frac{l_4 \mathcal{L}_{0.5}\binom{1}{1}^2 \mathcal{L}_1\binom{1;1}{1;1}}{24T^2 l_2^4} \\
&= -\frac{\mathcal{L}_1\binom{2}{1}^2 l_4}{24T^2 l_2^4} - \frac{l_4 \mathcal{L}_{0.5}\binom{1}{1}\mathcal{L}_{1.5}\binom{1;2}{1;1}}{24T^2 l_2^4} - \frac{l_4 \mathcal{L}_{0.5}\binom{1}{1}\mathcal{L}_{1.5}\binom{2;1}{1;1}}{24T^2 l_2^4} - \frac{l_4 \mathcal{L}_{0.5}\binom{1}{1}\mathcal{L}_{1.5}\binom{1;1;1}{1;1;1}}{24T^2 l_2^4} \\
&\quad - \frac{\mathcal{L}_1\binom{2}{1} l_4 \mathcal{L}_1\binom{1;1}{1;1}}{24T^2 l_2^4}
\end{aligned}
$$

$$= -\frac{l_4\mathcal{L}_2\binom{2;2}{1;1}}{12T^2l_2^4} - \frac{l_4\mathcal{L}_2\binom{1;1;2}{1;1;1}}{12T^2l_2^4} - \frac{l_4\mathcal{L}_2\binom{1;2;1}{1;1;1}}{12T^2l_2^4} - \frac{l_4\mathcal{L}_2\binom{2;1;1}{1;1;1}}{12T^2l_2^4} - \frac{l_4\mathcal{L}_2\binom{1;1;1;1}{1;1;1;1}}{24T^2l_2^4}$$

$$- \frac{\mathcal{L}_1\binom{2}{1}^2 l_4}{24T^2l_2^4} + O_p(T^{-5/2})$$

$$= -\frac{\mathcal{L}_3\binom{1;1;2;1}{1;1;1;4}}{4T^2l_2^4} - \frac{\mathcal{L}_3\binom{1;1;1;1;1}{1;1;1;1;4}}{24T^2l_2^4} - \frac{l_4\mathcal{L}_2\binom{2;2}{1;1}}{12T^2l_2^4} - \frac{\mathcal{L}_1\binom{2}{1}^2 l_4}{24T^2l_2^4} + O_p(T^{-5/2}),$$

which are the expressions given in Appendix B.1.

B.5 Intermediate steps in step 2

It can be derived that

$$\frac{l_1\mathcal{L}_1\binom{1,1}{1,2}}{Tl_2^2} = \frac{\mathcal{L}_1\binom{2,1}{1,2}}{T^2l_2^2} + \frac{\mathcal{L}_{1.5}\binom{1;1,1}{1;1,2}}{T^{1.5}l_2^2},$$

$$\frac{l_1^2\mathcal{L}_1\binom{2}{2}}{2Tl_2^3} = \frac{\mathcal{L}_2\binom{1;1;2}{1;1;2}}{2T^2l_2^3} + \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{2}{2}}{2T^2l_2^3} + O_p(T^{-5/2}),$$

$$\frac{l_1^2\mathcal{L}_1\binom{1,1}{1,3}}{2Tl_2^3} = \frac{\mathcal{L}_2\binom{1;1;1,1}{1;1;1,3}}{2T^2l_2^3} + \frac{\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{1,1}{1,3}}{2T^2l_2^3} + O_p(T^{-5/2}),$$

$$\frac{l_1^2l_4\mathcal{L}_1\binom{2}{1}}{4Tl_2^4} = \frac{\mathcal{L}_3\binom{1;1;2;1}{1;1;1;4}}{4T^2l_2^4} + \frac{l_4\mathcal{L}_1\binom{2}{1}^2}{4T^2l_2^4} + O_p(T^{-5/2}),$$

$$\frac{3l_1^2l_3\mathcal{L}_1\binom{1,1}{1,2}}{2Tl_2^4} = \frac{3\mathcal{L}_3\binom{1;1;1;1,1}{1;1;3;1,2}}{2T^2l_2^4} + \frac{3l_3\mathcal{L}_1\binom{2}{1}\mathcal{L}_1\binom{1,1}{1,2}}{2T^2l_2^4} + O_p(T^{-5/2}),$$

$$\frac{3l_1^2l_3^2\mathcal{L}_1\binom{2}{1}}{4Tl_2^5} = \frac{3\mathcal{L}_4\binom{1;1;2;1;1}{1;1;1;3;3}}{4T^2l_2^5} + \frac{3l_3^2\mathcal{L}_1\binom{2}{1}^2}{4T^2l_2^5} + O_p(T^{-5/2}),$$

$$\frac{l_1l_3\mathcal{L}_1\binom{2}{1}}{2Tl_2^3} = \frac{\mathcal{L}_2\binom{2;1,1}{1;1,3}}{2T^2l_2^3} + \frac{\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{2T^{1.5}l_2^3} + \frac{l_3\mathcal{L}_1\binom{3}{1}}{2T^2l_2^3} + O_p(T^{-5/2}),$$

where, using the expansion in (5.4),

$$\frac{\mathcal{L}_{1.5}\binom{1;1,1}{1;1,2}}{T^{1.5}l_2^2} = \frac{3\mathcal{L}_{1.5}\binom{1;1,1}{1;1,2}}{T^{1.5}\left(\mathbb{E}l_2\right)^2} - \frac{2\mathcal{L}_{2.5}\binom{1;1;1,1}{1;2;1,2}}{T^{1.5}\left(\mathbb{E}l_2\right)^3} - \frac{2\mathcal{L}_2\binom{1,1;1,1}{1,2;1,2}}{T^2\left(\mathbb{E}l_2\right)^3} + O_p(T^{-5/2}),$$

$$\frac{\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{2T^{1.5}l_2^3} = -\frac{3\mathcal{L}_{3.5}\binom{1;2;1;1}{1;1;3;2}}{2T^{1.5}\left(\mathbb{E}l_2\right)^4} + \frac{2\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{T^{1.5}\left(\mathbb{E}l_2\right)^3} - \frac{3\mathcal{L}_3\binom{2;1;1,1}{1;3;1,2}}{2T^2\left(\mathbb{E}l_2\right)^4} + O_p(T^{-5/2}).$$

Using Lemma 1, it follows that

$$\mathbb{E}\frac{3\mathcal{L}_{1.5}\binom{1;1,1}{1;1,2}}{T^{1.5}\left(\mathbb{E}l_2\right)^2} = \mathbb{E}\frac{2\mathcal{L}_{2.5}\binom{1;1;1,1}{1;2;1,2}}{T^{1.5}\left(\mathbb{E}l_2\right)^3} = \mathbb{E}\frac{3\mathcal{L}_{3.5}\binom{1;2;1;1}{1;1;3;2}}{2T^{1.5}\left(\mathbb{E}l_2\right)^4} = \mathbb{E}\frac{2\mathcal{L}_{2.5}\binom{1;2;1}{1;1;3}}{T^{1.5}\left(\mathbb{E}l_2\right)^3} = 0,$$

so the corresponding terms in the expansions can be dropped. This leads to (5.6) in Appendix B.2.

C Review of the derivation of Arellano and Hahn (2016)

We rederive the first-order correction ($k = 1$) of Arellano and Hahn (2016). In a series of remarks, we highlight the differences with the derivation of the second-order correction ($k = 2$).

Let $\widehat{l} := \widehat{l}(\theta)$ and $l := l(\theta)$. For a regular problem, $\widehat{l}$ can be Taylor-expanded around $\alpha_i(\theta)$, giving

$$\begin{aligned} \widehat{l} &= l + l_1 (\widehat{\alpha}_i(\theta) - \alpha_i(\theta)) + \frac{1}{2} l_2 (\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^2 + O_p(T^{-3/2}), \\ l &= \widehat{l} - l_1 (\widehat{\alpha}_i(\theta) - \alpha_i(\theta)) - \frac{1}{2} l_2 (\widehat{\alpha}_i(\theta) - \alpha_i(\theta))^2 + O_p(T^{-3/2}). \end{aligned} \tag{5.8}$$

Remark 9 *For $k = 2$, (5.8) needs to be extended so that the remainder term is $O_p(T^{-5/2})$. This is fairly straightforward.*

Similarly, $\widehat{l}_1 = 0$ can be Taylor-expanded around $\alpha_i(\theta)$, yielding

$$0 = l_1 + l_2 (\widehat{\alpha}_i(\theta) - \alpha_i(\theta)) + O_p(T^{-1}) \tag{5.9}$$

and, since $l_2 < 0$,

$$\widehat{\alpha}_i(\theta) - \alpha_i(\theta) = -\frac{l_1}{l_2} + O_p(T^{-1}). \tag{5.10}$$

Remark 10 *For $k = 2$, (5.9) has to be extended so that the remainder term is $O_p(T^{-2})$. From this, an asymptotic representation of $\widehat{\alpha}_i(\theta) - \alpha_i(\theta)$ can be derived whose remainder term is also $O_p(T^{-2})$. This is not difficult either; the technique is given in, e.g., Pace and Salvan (1997, chap. 9).*

Next, (5.8) and (5.10) are combined to give

$$\begin{aligned} l &= \widehat{l} - l_1 \left(-\frac{l_1}{l_2}\right) - \frac{1}{2} l_2 \left(-\frac{l_1}{l_2}\right)^2 + O_p(T^{-3/2}) \\ &= \widehat{l} + \frac{1}{2} \frac{l_1^2}{l_2} + O_p(T^{-3/2}). \end{aligned}$$

Remark 11 *For $k = 2$, the higher-order representation of $\widehat{\alpha}_i(\theta) - \alpha_i(\theta)$ and the higher-order version of (5.8) have to be combined. This requires raising a polynomial expression to the power 4. To manage the expression, the technique of Provost and Ratemi (2011) may be used.*

At this point, note that, as $\widehat{l}_1 = 0$, if one replaces l_1^2 with $\widehat{l}_1^2$, the ratio l_1^2/l_2 disappears completely. Therefore, a more refined estimator of l_1^2/l_2 needs to be constructed. By the definition of l_1,

$$\begin{aligned} l_1^2 = &\frac{1}{T^2} \sum_t \left(\nabla_{a_i} \log f(Y_{it}; \theta, a_i)\right)^2 \Big|_{a_i = \alpha_i(\theta)} \\ &+ \frac{1}{T^2} \sum_{t_1 \neq t_2} \nabla_{a_i} \log f(Y_{it_1}; \theta, a_i) \nabla_{a_i} \log f(Y_{it_2}; \theta, a_i) \Big|_{a_i = \alpha_i(\theta)}. \end{aligned}$$

By independence,

$$\mathbb{E}\sum_{t_1\neq t_2}\nabla_{a_i}\log f\left(Y_{it_1};\theta,a_i\right)\nabla_{a_i}\log f\left(Y_{it_2};\theta,a_i\right)\big|_{a_i=\alpha_i(\theta)}=0. \tag{5.11}$$

Hence

$$\mathbb{E}l=\mathbb{E}\widehat{l}+\frac{\mathbb{E}b_1}{T}+O(T^{-2})$$

where

$$b_1=\frac{1/T\sum_t\left(\nabla_{a_i}\log f\left(Y_{it};\theta,a_i\right)\right)^2\Big|_{a_i=\alpha_i(\theta)}}{2l_2}.$$

Remark 12 *For $k=2$, the identification of the terms with zero expectation, similar to (5.11), is necessary. This step is rather involved.*

Replacing b_1 with $\widehat{b}_1$ introduces a bias since, typically,

$$\mathbb{E}\widehat{b}_1=\mathbb{E}b_1+O(T^{-1}).$$

However, for $k=1$, this bias can be neglected because

$$\begin{aligned}\mathbb{E}l&=\mathbb{E}\widehat{l}+\frac{\mathbb{E}b_1}{T}+O(T^{-2})\\&=\mathbb{E}\widehat{l}+\frac{\mathbb{E}\widehat{b}_1+O(T^{-1})}{T}+O(T^{-2})\\&=\mathbb{E}\widehat{l}+\frac{\mathbb{E}\widehat{b}_1}{T}+O(T^{-2}).\end{aligned}$$

Therefore, the first-order corrected log-likelihood follows as

$$\widehat{l}^{(1)}\left(\theta\right):=\widehat{l}+\frac{\widehat{b}_1}{T}.$$

$\widehat{l}^{(1)}\left(\theta\right)$ can be constructed from the sample since $\widehat{l}=\widehat{l}\left(\theta\right)$ and $\widehat{b}_1=\widehat{b}_1\left(\theta\right)$ depend only on known quantities, $\widehat{\alpha}_i\left(\theta\right)$ and Y_{it}.

Remark 13 *For $k=2$, the bias introduced by replacing b_1 with $\widehat{b}_1$ must also be taken into account. To deal with this, a procedure similar to that dealing with the pair $(l,\widehat{l})$ has to be applied to the pair $(b_1,\widehat{b}_1)$. This step is also involved.*

D Expansion of fixed effect

Proof (Proposition 2). As in Cox and Snell (1968), a Taylor expansion of $\widehat{l}_1=0$ around $\alpha_i\left(\theta\right)$ gives

$$0=l_1+l_2\left(\widehat{\alpha}_i\left(\theta\right)-\alpha_i\left(\theta\right)\right)+\frac{1}{2!}l_3\left(\widehat{\alpha}_i\left(\theta\right)-\alpha_i\left(\theta\right)\right)^2$$

$$+\frac{1}{3!}l_4\left(\widehat{\alpha}_i(\theta)-\alpha_i(\theta)\right)^3+O_p(T^{-2})$$

and, on rearranging,

$$\widehat{\alpha}_i(\theta)-\alpha_i(\theta)=-\frac{l_1}{l_2}-\frac{1}{2!}\frac{l_3}{l_2}\left(\widehat{\alpha}_i(\theta)-\alpha_i(\theta)\right)^2 -\frac{1}{3!}\frac{l_4}{l_2}\left(\widehat{\alpha}_i(\theta)-\alpha_i(\theta)\right)^3+O_p(T^{-2}). \tag{5.12}$$

We now seek a representation

$$\widehat{\alpha}_i(\theta)-\alpha_i(\theta)=a_{1/2}+a_{2/2}+a_{3/2}+O_p(T^{-2}) \tag{5.13}$$

where the $a_{j/2}$ are random variables satisfying $a_{j/2}=O_p(T^{-j/2})$ and $a_{j/2}\neq o_p(T^{-j/2})$. The $a_{j/2}$ can be solved via a recursive procedure similar to the one described in Pace and Salvan (1997, chap. 9). In particular, combining (5.12) and (5.13) gives

$$\begin{aligned} &a_{1/2}+a_{2/2}+a_{3/2}\\ &\quad=-\frac{l_1}{l_2}-\frac{1}{2!}\frac{l_3}{l_2}(a_{1/2}+a_{2/2}+a_{3/2})^2-\frac{1}{3!}\frac{l_4}{l_2}(a_{1/2}+a_{2/2}+a_{3/2})^3+O_p(T^{-2})\\ &\quad=-\frac{l_1}{l_2}-\frac{1}{2!}\frac{l_3}{l_2}(a_{1/2}^2+2a_{1/2}a_{2/2})-\frac{1}{3!}\frac{l_4}{l_2}a_{1/2}^3+O_p(T^{-2}).\end{aligned}$$

On grouping the terms by their stochastic order,

$$a_{1/2}=-\frac{l_1}{l_2},\qquad a_{2/2}=-\frac{l_3}{2l_2}a_{1/2}^2,\qquad a_{3/2}=-\frac{l_3}{l_2}a_{1/2}a_{2/2}-\frac{1}{6}\frac{l_4}{l_2}a_{1/2}^3,$$

and, after a recursive substitution,

$$a_{1/2}=-\frac{l_1}{l_2},\qquad a_{2/2}=-\frac{l_1^2l_3}{2l_2^3},\qquad a_{3/2}=-\frac{l_3^2l_1^3}{2l_2^5}+\frac{l_1^3l_4}{6l_2^4}.$$

This gives the result in Proposition 2. ∎

Remark 14 (Stochastic order) *The order of each term can be easily identified. As* $l_1=O_p(T^{-1/2})$ *and* $l_r=O_p(1)$ *for* $1<r\leq 3$*, it is clear that* $a_{1/2}=O_p(T^{-1/2})$*,* $a_{2/2}=O_p(T^{-1})$*, and* $a_{3/2}=O_p(T^{-3/2})$*; and that, when* $l_1\neq o_p(T^{-1/2})$ *and* $l_r\neq o_p(1)$ *for* $1<r\leq 3$*,* $a_{1/2}\neq o_p(T^{-1/2})$*,* $a_{2/2}\neq o_p(T^{-1})$*, and* $a_{3/2}\neq o_p(T^{-3/2})$*.*

E Higher-order expansion of fixed effect

Suppose Assumptions 1 and 2 hold with r up to 9 in Assumption 1. Then

$$\widehat{\alpha}_i(\theta)-\alpha_i(\theta)=\sum_{j=1}^{8}a_{j/2}+O_p(T^{-9/2}),$$

where $a_{1/2}$ to $a_{3/2}$ are given in Appendix D and

$$a_{4/2}=\frac{5l_1^4l_3l_4}{12l_2^6}-\frac{5l_1^4l_3^3}{8l_2^7}-\frac{l_1^4l_5}{24l_2^5},$$

$$
\begin{aligned}
a_{5/2} &= \frac{l_1^5 l_6}{120 l_2^6} - \frac{l_1^5 l_4^2}{12 l_2^7} - \frac{7 l_1^5 l_3^4}{8 l_2^9} - \frac{l_1^5 l_3 l_5}{8 l_2^7} + \frac{7 l_1^5 l_3^2 l_4}{8 l_2^8}, \\
a_{6/2} &= \frac{7 l_1^6 l_3 l_6}{240 l_2^8} - \frac{21 l_1^6 l_3^5}{16 l_2^{11}} - \frac{l_1^6 l_7}{720 l_2^7} + \frac{7 l_1^6 l_4 l_5}{144 l_2^8} - \frac{7 l_1^6 l_3 l_4^2}{18 l_2^9} - \frac{7 l_1^6 l_3^2 l_5}{24 l_2^9} + \frac{7 l_1^6 l_3^3 l_4}{4 l_2^{10}}, \\
a_{7/2} &= \frac{l_1^7 l_8}{5040 l_2^8} - \frac{l_1^7 l_5^2}{144 l_2^9} + \frac{l_1^7 l_4^3}{18 l_2^{10}} - \frac{33 l_1^7 l_3^6}{16 l_2^{13}} - \frac{5 l_1^7 l_3^2 l_4^2}{4 l_2^{11}} - \frac{l_1^7 l_3 l_7}{180 l_2^9} - \frac{l_1^7 l_4 l_6}{90 l_2^9} \\
&\quad + \frac{3 l_1^7 l_3^2 l_6}{40 l_2^{10}} - \frac{5 l_1^7 l_3^3 l_5}{8 l_2^{11}} + \frac{55 l_1^7 l_3^4 l_4}{16 l_2^{12}} + \frac{l_1^7 l_3 l_4 l_5}{4 l_2^{10}}, \\
a_{8/2} &= \frac{l_1^8 l_3 l_8}{1120 l_2^{10}} - \frac{429 l_1^8 l_3^7}{128 l_2^{15}} - \frac{55 l_1^8 l_3^3 l_4^2}{16 l_2^{13}} - \frac{l_1^8 l_9}{40320 l_2^9} + \frac{l_1^8 l_4 l_7}{480 l_2^{10}} + \frac{l_1^8 l_5 l_6}{320 l_2^{10}} \\
&\quad - \frac{5 l_1^8 l_3 l_5^2}{128 l_2^{11}} - \frac{5 l_1^8 l_4^2 l_5}{96 l_2^{11}} + \frac{55 l_1^8 l_3 l_4^3}{144 l_2^{12}} - \frac{l_1^8 l_3^2 l_7}{64 l_2^{11}} + \frac{11 l_1^8 l_3^3 l_6}{64 l_2^{12}} - \frac{165 l_1^8 l_3^4 l_5}{128 l_2^{13}} \\
&\quad + \frac{429 l_1^8 l_3^5 l_4}{64 l_2^{14}} + \frac{55 l_1^8 l_3^2 l_4 l_5}{64 l_2^{12}} - \frac{l_1^8 l_3 l_4 l_6}{16 l_2^{11}}.
\end{aligned}
$$

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological) 32*(2), 283–301.

Arellano, M. and J. Hahn (2016). A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. *Global Economic Review 45*(3), 251–274.

Bartlett, M. S. (1953a). Approximate confidence intervals. *Biometrika 40*(1/2), 12–19.

Bartlett, M. S. (1953b). Approximate confidence intervals. II. more than one unknown parameter. *Biometrika 40*(3/4), 306–317.

Blundell, R., R. Griffith, and F. Windmeijer (2002). Individual effects and dynamics in count data models. *Journal of Econometrics 108*(1), 113–131.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies 47*(1), 225–238.

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

Cox, D. (1958). Two further applications of a model for binary regression. *Biometrika 45*(3/4), 562–565.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological) 30*(2), 248–275.

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies 82*(3), 991–1030.

Fernández-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics 150*(1), 71–85.

Ferrari, S. L., D. A. Botter, G. M. Cordeiro, and F. Cribari-Neto (1996). Second-and third-order bias reduction for one-parameter family models. *Statistics & probability letters 30*(4), 339–345.

Greene, W., C. Han, and P. Schmidt (2002). The bias of the fixed effects estimator in nonlinear models. *Unpublished manuscript*.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica 70*(4), 1639–1657.

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

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

Haldane, J. and S. M. Smith (1956). The sampling distribution of a maximum-likelihood estimate. *Biometrika 43*(1/2), 96–103.

Heckman, J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data*. MIT Press.

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

Lancaster, T. (2002). Orthogonal parameters and panel data. *The Review of Economic Studies 69*(3), 647–666.

Newey, W. K. and D. McFadden (1994). *Handbook of Econometrics*, Volume 6A. Elsevier.

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

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*(6), 1417–1426.

Pace, L. and A. Salvan (1997). *Principles of statistical inference: from a Neo-Fisherian perspective*, Volume 4. World Scientific.

Provost, S. B. and W. M. Ratemi (2011). Polynomial expansions via embedded Pascal¡¯s triangles. *Acta et Commentationes Universitatis Tartuensis de Mathematica 15*(1), 45–60.

Rice, S. H. (2008). A stochastic version of the price equation reveals the interplay of deterministic and stochastic processes in evolution. *BMC evolutionary biology 8*(1), 262.

Rilstone, P., V. Srivastava, and A. Ullah (1996). The second-order bias and mean squared error of nonlinear estimators. *Journal of Econometrics 75*(2), 369–395.