Econometric Reviews

Taylor & Francis
Taylor & Francis Group

ISSN: 0747-4938 (Print) 1532-4168 (Online) Journal homepage: www.tandfonline.com/journals/lecr20

Nuisance parameters, modified profile likelihood and Jacobian prior

Guangjie Li & Roberto Leon-Gonzalez

ECONOMETRIC REVIEWS
https://doi.org/10.1080/07474938.2025.2581301

Taylor & Francis
Taylor & Francis Group

Nuisance parameters, modified profile likelihood and Jacobian prior

Guangjie Li[a] and Roberto Leon-Gonzalez[b]

[a]Cardiff Business School, Cardiff University, Cardiff, Wales; [b]Fellow of the Rimini Centre for Economic Analysis, National Graduate Institute for Policy Studies, Japan

ABSTRACT

In a model with nuisance parameters, the maximum likelihood estimators (MLE) of the parameters of interest can be biased. One can reduce the bias due to the presence of the nuisance parameters by removing the $O(1)$ bias of the profile likelihood score. To achieve this, we propose the Jacobian integrated likelihood (JIL) obtained by using a prior consisting of the Jacobian determinant of the new nuisance parameters, which are functions of the original nuisance parameters and are independent of the dependent variable. Our JIL is closely related to the modified profile likelihood (MPL) in Barndorff-Nielsen and Cox (1994). We propose the adjusted MPL, which is easier to compute and can also remove the $O(1)$ bias of the profile likelihood score. For panel fixed effects models, both the JIL and the adjusted MPL can remove the bias of order $O(T^{-1})$ in the MLE as the cross-sectional size (N) increases. We give the conditions when the estimators from the adjusted MPL and the JIL are the same and consistent with $T = o(N)$. Although the adjusted MPL and the JIL do not always exist, one can use their first-order conditions to obtain bias-reduced estimators. The theoretical results are demonstrated by panel binary choice models and dynamic panel linear models with exogenous and predetermined regressors.

ARTICLE HISTORY
Received 23 May 2024
Accepted 22 October 2025

KEYWORDS
Bayesian estimation; incidental parameter; maximum likelihood; nuisance parameter panel fixed effects models

JEL CLASSIFICATION CODE:
C11; C13; C15

1. Introduction

It is well known in statistics that maximum likelihood estimators (MLE) are in general biased, see for example 5.3 in Barndorff-Nielsen and Cox (1994). The parameters in a likelihood function could be classified as either nuisance parameters or parameters of interest. The MLE bias in the parameters of interest could exist even without any nuisance parameters in the model while the presence of the nuisance parameters could add another source of bias to the parameters of interest. Using the expansion method of Barndorff-Nielsen and Cox (1994), we derive the $O(T^{-1})$ bias of the parameters of interest, where T is the time-series sample size, with the presence of nuisance parameters. We find that the nuisance parameters affect the $O(T^{-1})$ MLE bias of the parameters of interest through the profile likelihood scores and their cross elements with the parameters of interest in the information matrix. To remove the bias, we extend the score correction method of Firth (1993) to models with nuisance parameters. In comparison to the one-step and iterated bias correction methods used in Hahn and Newey (2004), Hahn and Kuersteiner (2011) and Fernandez-Val (2009), with our method one does not need to first obtain the MLE and our estimator is also free of the $O(T^{-1})$ bias.

For panel fixed effects models, the number of the fixed-effect parameters, which capture the heterogeneity of the economic agents, increases with the cross-sectional sample size (N). They are called incidental parameters by Neyman and Scott (1948), while the parameters with fixed dimensions are called

CONTACT Guangjie LI ligj@cardiff.ac.uk Cardiff Business School, Cardiff University, Cardiff, Wales.

common parameters. For a fixed time-series sample size, the MLE of the common parameters can be inconsistent as N grows to infinity, which is the incidental parameter problem and has been extensively discussed in econometrics and statistics literature: see for example Heckman (1981), Lancaster (2000), Greene (2004), Arellano and Hahn (2007), Bester and Hansen (2009) and Moreira (2009). In this paper, we treat the fixed-effect parameters as nuisance parameters and the common parameters as parameters of interest. One can obtain the estimators for the common parameters with bias up to $O(T^{-2})$ for large N if the $O(1)$ bias in the average profile likelihood scores for the common parameters is removed. To achieve this, we propose the Jacobian integrated likelihood (JIL) obtained by using a prior equal to the Jacobian determinant of the new incidental parameters, which are functions of the original incidental parameters. Our work is related to Arellano and Bonhomme (2009), who studied the incidental parameter problem from a Bayesian perspective, though their prior depends on the dependent variable or the true values of the model parameters, while ours does not. Our approach is an extension of the information orthogonal reparameterization method used by Lancaster (2002), who tried to reparameterize the likelihood function such that the new incidental parameters are information orthogonal to the common parameters. We show that information orthogonal reparameterization is not necessary to produce bias-reducing priors. We propose the concept of weak information orthogonality. When the incidental parameters are weakly information orthogonal to the common parameters, the Jacobian prior can be flat and is bias-reducing. We also show that our Jacobian prior can be viewed as the integrating factor for the differential equation used to obtain the information orthogonal reparameterization. The Jacobian prior can exist even when it is impossible to find the information orthogonal reparameterization, for example, for the linear dynamic panel model of order p ($p \geq 1$) with exogenous regressors. The JIL is closely related to the modified profile likelihood (MPL) in Barndorff-Nielsen and Cox (1994), which exists in theory, but is hard to be calculated in practice. We propose a computationally easier MPL, the adjusted MPL, which can also remove the profile likelihood score bias asymptotically. We give the conditions under which the adjusted MPL and the JIL are the same and consistent and show that the linear dynamic panel model is such an example. Though even the Jacobian prior and the adjusted MPL do not always exist, as in the linear dynamic panel model with predetermined regressors, one can use the first order conditions from the JIL and the adjusted MPL to obtain consistent estimators. For nonlinear panel models, such as logit and probit models, the bias-reduced estimators from the JIL and the adjusted MPL are in general different for finite samples.

The plan of the article is as follows. Section 2 discusses the models with nuisance parameters in general and how the JIL and the adjusted MPL are constructed. Section 3 studies how the JIL and the adjusted MPL can be applied to panel fixed effects models. Section 3.1 and 3.2 show the examples for the theoretical results along with Monte Carlo evidence to support our results before Section 4 concludes.

2. Models with nuisance parameters in general

We denote the likelihood function by $p(Y|\eta, X)$ with $\eta = (\theta, f)$. Y is the collection of the individual observations of the dependent variable and X is the collection of the explanatory variables and can include the initial observations of Y. The parameters to be estimated are put into two categories: the parameters of interest, denoted by θ, and the nuisance parameters, f. In this section, we mainly treat θ as a scalar, though the results below can be extended to the case when θ is a vector. $l(\theta, f)$ or $l(\eta)$ denotes the log likelihood function, $\ln p(Y|\eta, X)$. We will use $r, s, .., v, w$ to index the elements in f and a^r or a_r denotes the r-th element of vector a. The Einstein summation convention is used here: for instance, $a^i b_i = \sum_i a_i b_i$. Suppose $R_1 = r_1 \ldots r_m$ and $R_2 = u_1 \ldots u_m$, which are arbitrary index sets, then l_{R_1} denotes the derivative of the log likelihood function with respect to the elements of f indicated by R_1. For example, if $R_1 = r_1 r_2$, $l_{R_1} = \frac{\partial^2 l(\theta, f)}{\partial f^{r_1} \partial f^{r_2}}$. Similarly, l_θ is the partial derivative with respect to θ. Additionally, $V_{R_1} = E(l_{R_1}) = \int l_{R_1} p(Y|\eta, X) dY$, $V_{R_1,R2} = E(l_{R_1} l_{R_2})$ and $H_{R_1} = l_{R_1} - V_{R_1}$. I_{rs} denotes the (r,s)-component of the matrix $I_{ff} = E(-l_{ff}) = E(-\frac{\partial^2 l(\theta,f)}{\partial f \partial f'})$ while I^{rs} is the (r,s)-element of I_{ff}^{-1} and $V^{R_1} = I^{r_1u_1} \dots I^{r_mu_m} V_{R_2}$. Throughout the paper, $|\cdot|$ is the operation to obtain the absolute value of the determinant of a square matrix and our results will not rely on signed determinants.

The following are the assumptions we use to derive our theoretical results:

Assumption 1. The dependent variable generated by the likelihood function evaluated at the unique true value of η, y_t $(t = 1, 2, \dots, T)$ and the explanatory covariates, x_t, which the likelihood function is conditional on and can include the past values of y_t, are strictly stationary, finite order Markov processes, which are Harris recurrent and aperiodic.

Assumption 2. (i) The summands of $l(\eta) = \sum_{t=1}^{T} \ln p(y_t|\eta, x_t)$ consist of the same function and differ only in y_t and x_t. (ii) $l(\eta)$ is five times continuously differentiable with respect to any elements of η in a neighborhood of the true. (iii) $E(|\frac{\partial^4 \ln p(y_t|\eta,x_t)}{\partial\eta^{v_1}\partial\eta^{v_2}\partial\eta^{v_3}\partial\eta^{v_4}}|^{4+\gamma}) < \infty$ for $\gamma > 0$ and the element index $v_i = 0, 1, 2, \dots$ for $i, j = 1, \dots, 4$. If $v_i = 0$, the derivative order will be reduced, e.g. $\frac{\partial \ln p(y_t|\eta,x_t)}{\partial\eta^{v_i}} = \ln p(y_t|\eta, x_t)$. (iv) $E(|\frac{\partial^5 \ln p(y_t|\theta,f,x_t)}{\partial\eta^{v_1}\partial\eta^{v_2}\partial\eta^{v_3}\partial\eta^{v_4}\partial\eta^{v_5}}|) < \infty$. (v) The diagonal elements of $E(\frac{\partial l(\eta)}{\partial\eta}\frac{\partial l(\eta)}{\partial\eta'})$ (a positive definite matrix) will tend to infinity as T grows.

Assumption 3. The interchange of the operations of differentiation with respect to η and integration of $\int p(Y|\eta, X)dY = 1$ discussed in 17.16 of Ord, Arnold, and Stuart (1999) are assumed to be valid such that Bartlett's identities, see for example 5.2 in Barndorff-Nielsen and Cox (1994), hold up to the third order.

Assumption 4. Denote $Z_t = (y_t, x_{t+1})$. The sequence $\{Z_t\}$ is defined on a probability space $(\Omega, \mathcal{F}, P)$ with the mixing coefficient: $\alpha_Z(n) = \sup|P(A \cap B) - P(A)P(B)|$, where $A \in \mathcal{F}_{-\infty}^{t}$, $B \in \mathcal{F}_{t+n}^{\infty}$ and $\mathcal{F}_j^i$ denotes the σ-field generated by the random variables Z_t with $j \le t \le i$ $(t \in \mathbb{Z})$, such that $\sum_{n=1}^{\infty} n[\alpha_Z(n)]^{\frac{\gamma}{3(3+\gamma)}} < \infty$, where γ appears in Assumption 2.

Under Assumption 1, $\{Z_t\}$ is strong mixing $(\alpha_Z(n) \to 0)$ according to Corollary 3.6 in Bradley (2005). Assumption 2 and 4, which further restricts the rate of convergence, $\alpha_Z(n)$, are required to derive the results in the subsequent sections. In contrast to the existing literature, which requires the data to be exponentially mixing, e.g. Hahn and Kuersteiner (2011), we only assumes normal mixing, which could potentially cover more models. If one just needs to prove Theorem 2.1 below, one can relax the moment conditions in Assumption 2 to $E(|\frac{\partial \ln p(y_t|\eta,x_t)}{\partial\eta^{v_1}}|^2)$, $E(|\frac{\partial^2 \ln p(y_t|\eta,x_t)}{\partial\eta^{v_1}\partial\eta^{v_2}}|)$ and $E(|\frac{\partial^3 \ln p(y_t|\eta,x_t)}{\partial\eta^{v_1}\partial\eta^{v_2}\partial\eta^{v_3}}|)$ being finite.

Theorem 2.1. *Under Assumption 1 to 3, the model satisfies the general stability conditions of (3.11) in Barndorff-Nielsen and Cox (1994): l_r and l_θ are of order $O_p(T^{\frac{1}{2}})$ and the MLE of η, denoted as $\hat{\eta} = (\hat{\theta}, \hat{f})$ satisfies $\hat{\eta} - \eta = O_p(T^{-\frac{1}{2}})$.*

2.1. MLE Bias in the parameters of interest

Denote the MLE of f for a given θ as $\hat{f}_{(\theta)}$. We assume Assumption 1 to 4 hold in this subsection. The properties of the profile likelihood score can be described in the lemma below.

Lemma 2.1. *The total derivative of the profile likelihood, $l(\theta, \hat{f}_{(\theta)})$, with respect to θ can have the following asymptotic expansion:*

$$\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta} = l_\theta - I_{\theta r}I^{rs}l_s \blacktriangledown + H_{\theta r}I^{rs}l_s - \frac{1}{2}I_{\theta r}I^{ru_1}I^{su_2}I^{vu_3}V_{u_1u_2u_3}l_sl_v \\ - I_{\theta r}I^{rs}I^{vu}H_{sv}l_u + \frac{1}{2}V_{\theta rs}I^{ru}I^{sv}l_ul_v \blacktriangledown + O_p(T^{-\frac{1}{2}}). \tag{1}$$

where $I_{\theta r} = E(-\frac{\partial^2 l(\theta,f)}{\partial\theta\partial f^r})$ *and the symbol* $\blacktriangledown$ *indicates a change in asymptotic magnitude of order. All the derivatives of the log likelihood function are evaluated at the true* θ *and* f. *The expectation of the profile likelihood score is*

$$E\left[\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}\right] = B(\theta,f) \blacktriangledown + O(T^{-1}) \tag{2}$$

$$= \frac{1}{2}I^{rs}\left(2V_{\theta r,s} + V_{\theta rs}\right) - \frac{1}{2}I^{rs}\left(V_{rsv} + 2V_{rv,s}\right)I^{vu}I_{u\theta} \blacktriangledown + O(T^{-1}) \tag{3}$$

$$= -\frac{1}{2}I^{rs}\left(V_{\theta,rs} + V_{\theta,r,s}\right) + \frac{1}{2}I^{rs}\left(V_{r,s,v} + V_{rs,v}\right)I^{vu}I_{u\theta} \blacktriangledown + O(T^{-1}) \tag{4}$$

Note that the expectation in (2) is taken after dropping the $o_p(T^{-1})$ terms in the profile likelihood. Unlike the score of the likelihood function with zero expectation, the leading term $B(\theta,f)$ is of order $O(1)$. Equations (3) and (4) show that the bias of the profile likelihood score comes from two sources (see Appendix A.2): the bias in $\hat{f}_{(\theta)}$ ($\frac{1}{2}I^{rs}\left(V_{rsv} + 2V_{rv,s}\right)I^{vu}$), and the relationship between θ and f, which is captured by $I_{u\theta}$ and $2V_{\theta r,s} + V_{\theta rs}$ (or $V_{\theta,rs} + V_{\theta,r,s}$). Equation (3) is analogical to (12) under panel fixed effects models in Carro (2007) when the nuisance parameter is a scalar while Eq. (4) is the same as (8.61) in Barndorff-Nielsen and Cox (1994). Given our assumptions, the asymptotic expansion and the bias of $\hat{\theta}$ can be shown below.

Theorem 2.2. *Denote* $I^{\theta\theta} = (I_{\theta\theta} - I_{\theta r}I^{rs}I_{s\theta})^{-1}$, *the following holds:*

$$\hat{\theta} - \theta = I^{\theta\theta}\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta} + (I^{\theta\theta})^2R(\theta,f)(l_\theta - I_{\theta r}I^{rs}l_s) \\ + \frac{(I^{\theta\theta})^3}{2}W_{\theta\theta\theta}(l_\theta - I_{\theta r}I^{rs}l_s)^2 + O_p(T^{-\frac{3}{2}}) \tag{5}$$

$$E\left(\hat{\theta} - \theta\right) = b(\theta,f) \blacktriangledown + O(T^{-2}), \tag{6}$$

$$= I^{\theta\theta}B(\theta,f) + (I^{\theta\theta})^2C(\theta,f) \blacktriangledown + O(T^{-2}). \tag{7}$$

where

$$R(\theta,f) = H_{\theta\theta} - 2H_{\theta r}I^{rs}I_{s\theta} + I_{\theta r}I^{ru}H_{uv}I^{vs}I_{s\theta} \\ + \left(V_{\theta\theta r}I^{rs} - 2I^{rs}V_{\theta ru}I^{uv}I_{v\theta} + I_{\theta r}V^{rsv}I_{v\theta}\right)l_s = O_p(T^{\frac{1}{2}})$$
$$W_{\theta\theta\theta} = V_{\theta\theta\theta} - 3V_{\theta\theta r}I^{rs}I_{s\theta} + 3I_{\theta r}I^{rs}V_{sv\theta}I^{vu}I_{u\theta} - I_{\theta r}V^{rsv}I_{s\theta}I_{v\theta} = O(T)$$
$$C(\theta,f) = V_{\theta\theta,\theta} - I_{\theta r}I^{rs}V_{\theta\theta,s} - 2I^{rs}I_{s\theta}(V_{\theta r,\theta} - V_{\theta r,u}I_{\theta v}I^{vu}) \\ + I_{\theta r}I^{rs}I^{vu}I_{u\theta}(V_{sv,\theta} - I_{\theta v}I^{vw}V_{sv,w}) + \frac{1}{2}W_{\theta\theta\theta} = O(T) \tag{8}$$

The leading term of the bias ($b(\theta,f)$) comes from four sources: the bias of the profile likelihood score, the inter-dependence between the likelihood scores and the second order derivatives ($Cov(R(\theta,f), l_\theta - I_{\theta r}I^{rs}l_s)$), the mean of the third order derivatives ($W_{\theta\theta\theta}$) and the second moments of the likelihood scores ($E[(l_\theta - I_{\theta r}I^{rs}l_s)^2]$). The bias of the profile likelihood score as shown in (2) arises from the presence of the nuisance parameters while the bias from the other sources is partly inherent in the model and partly due to the relationship between the nuisance parameters and the parameters of interest. Note that when θ and f are **information orthogonal**: $I_{\theta f} = 0$, the second term in (7) will be reduced to

$(I_{\theta\theta})^{-2}\left(V_{\theta\theta,\theta}+\frac{1}{2}V_{\theta\theta\theta}\right)$, which may exist even without any nuisance parameters. The next lemma shows how to remove the $O(T^{-1})$ bias.

Lemma 2.2. *If $\tilde{\theta}$ is the solution for Eq. (9) below in a neighborhood which contains the true value, then (10) and (11) will hold.*

$$\left.\left[\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}-\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})\right]\right|_{\theta=\tilde{\theta}}=0, \tag{9}$$

$$\begin{aligned}\tilde{\theta}-\theta &= I^{\theta\theta}\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}-b(\theta,f)+(I^{\theta\theta})^2R(\theta,f)(l_\theta-I_{\theta r}I^{rs}l_s)\\ &\quad+\frac{(I^{\theta\theta})^3}{2}W_{\theta\theta\theta}(l_\theta-I_{\theta r}I^{rs}l_s)^2+O_p(T^{-\frac{3}{2}}) \end{aligned}\tag{10}$$

$$E\left(\tilde{\theta}-\theta\right)=O(T^{-2}). \tag{11}$$

Lemma 2.2 is an extension of the result in Firth (1993) to models with nuisance parameters. One can also use the one-step bias corrected estimator $\bar{\theta}^{(1)}=\hat{\theta}-b(\hat{\theta},\hat{f})$ or the iterated bias corrected estimator by solving $\bar{\theta}^{(\infty)}=\hat{\theta}-b\left(\bar{\theta}^{(\infty)},\hat{f}_{\bar{\theta}^{(\infty)}}\right)$ as discussed in Hahn and Newey (2004) and Fernandez-Val (2009), which are also biased up to $O(T^{-2})$.

Example 2.1. Consider a simple stable autoregressive model of order 1: $y_t=f+\rho y_{t-1}+\epsilon_t$ with $|\rho|<1$, where σ^2 is known with $\epsilon_t\sim i.i.d.N(0,\sigma^2)$. Suppose the parameter of interest is ρ while the nuisance parameter is f. Using the formula in (7), one can find that $b(\rho,f)=b(\rho)$ is composed of two components: the bias from the profile likelihood score due to the nuisance parameter is $-\frac{1+\rho}{T}$ while the bias from other sources is $-\frac{2\rho}{T}$. Note that even if σ^2 is treated as unknown and as another nuisance parameter, the results will not change. The $O(T^{-1})$ bias for this model, $-\frac{1+3\rho}{T}$, was previously found by Tanaka (1983), whose method is based on the Edgeworth expansion procedures. One can use (9) to obtain $\tilde{\rho}$ by solving

$$\frac{\sum_{t=1}^T(y_t-\bar{y})(y_{t-1}-\bar{y}\_)-\rho\sum_{t=1}^T(y_{t-1}-\bar{y}\_)^2}{\sigma^2}+\frac{1+3\rho}{1-\rho^2}=0. \tag{12}$$

where $\bar{y}=\frac{\sum y_t}{T}$ and $\bar{y}\_=\frac{\sum y_{t-1}}{T}$. When σ^2 is unknown, one can replace it by $\frac{\sum(y_t-\bar{y}-\rho(y_{t-1}-\bar{y}\_))^2}{T}$. One can expand (12) as a cubic equation, which can have multiple roots. Only one root should be chosen. For large T, the chosen root should be close to $\hat{\rho}=\frac{\sum(y_t-\bar{y})(y_{t-1}-\bar{y}\_)}{\sum(y_{t-1}-\bar{y}\_)^2}$. The one-step bias corrected estimator is $\bar{\rho}^{(1)}=\frac{T+3}{T}\hat{\rho}+\frac{1}{T}$ while the iterated bias corrected estimator is $\bar{\rho}^{(\infty)}=\frac{T}{T-3}\hat{\rho}+\frac{1}{T-3}$ with $T>3$. In comparison to $\hat{\rho}$, the $O(T^{-2})$ bias of $\bar{\rho}^{(1)}$ is that of $\hat{\rho}$ minus $\frac{3(1+3\rho)}{T^2}$, while the $O(T^{-2})$ bias of $\bar{\rho}^{(\infty)}$ is the same as that of $\hat{\rho}$.

2.2. Correction of the profile likelihood score

It could be complicated to calculate all the terms of the $O(T^{-1})$ bias in (6). A simpler way, which, though, does not necessarily remove all the $O(T^{-1})$ bias, would be to just correct the profile likelihood score by solving the equation below instead of (9) for θ,

$$\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}-B(\theta,\hat{f}_{(\theta)})=0, \tag{13}$$

where $B(\theta,f)$ is defined in (2). This method is especially relevant when the number of the nuisance parameters is large as in the case of the panel fixed effects models in Section 3, where removing the score bias can produce estimators free of the $O(T^{-1})$ bias. One can replace $B(\theta,\hat{f}_{(\theta)})$ with any functions of θ

whose expected difference from $B(\theta,f)$ is $O(T^{-1})$. For example, one can adapt the moment condition by setting Equation (3) in Woutersen (2003) equal to 0 and drop the lower order term to obtain

$$\left(l_\theta + \frac{I^{vu}}{2}\left[l_\theta - I_{\theta r} I^{rs} l_s \right]_{/vu} \right)\bigg|_{f=\hat{f}_{(\theta)}} = 0.$$

where $(a)_{/b}$ is the partial derivative of a with respect to b. Note that $E\{-\frac{I^{vu}}{2}\left[l_\theta - I_{\theta r} I^{rs} l_s \right]_{/vu} - B(\theta,f)\} = O(T^{-1})$ and I_{ff}^{-1} can be replaced with $(-l_{ff})^{-1}$.

The aim here is to find the likelihood function which produces the bias-reduced score. Two related methods will be discussed. The first method is the modified profile likelihood (MPL) described in Barndorff-Nielsen and Cox (1994, 8.1), see also Severini (2000, 9.3). The log MPL can be written as

$$l_{MP}(\theta) = -\frac{1}{2}\ln\left| -l_{ff}(\theta,\hat{f}_{(\theta)}) \right| + \ln D(\theta) + l(\theta,\hat{f}_{(\theta)}) \tag{14}$$

which is Equation (8.25) in Barndorff-Nielsen and Cox (1994) and where

$$D(\theta) = \frac{\left| -l_{ff}(\theta,\hat{f}_{(\theta)}) \right|}{\left| l_{f;\hat{f}}(\theta,\hat{f}_{(\theta)}) \right|} = \left| \frac{\partial \hat{f}_{(\theta)}}{\partial \hat{f}} \right|^{-1}. \tag{15}$$

$-l_{ff}(\theta,\hat{f}_{(\theta)})$ is the observed information matrix with f evaluated at $\hat{f}_{(\theta)}$ and $l_{f;\hat{f}} = \frac{\partial^2 l(\theta,f;\hat{\theta},\hat{f},a)}{\partial f \partial \hat{f}}$ is the mixed log model derivative defined in Barndorff-Nielsen and Cox (1994, 5.2), which is the second order derivative of the log likelihood function with respect to f and $\hat{f}$. The difficulty to use (15) is one has to write the likelihood function solely in terms of the parameters, their MLE and possibly the ancillary statistics (a) to obtain $l(\theta,f;\hat{\theta},\hat{f},a)$. In general the MLE do not always have closed forms and it could be difficult to find the ancillary statistics. Severini (2000) proposed the approximation $l_{f;\hat{f}}(\theta,\hat{f}_{(\theta)}) = I_{f,f}(\theta,\hat{f}_{(\theta)};\hat{\theta},\hat{f}) + O_p(T^{\frac{1}{2}})$ (9.5.4 in his book), where

$$I_{f,f}(\theta,f;\theta_0,f_0) = \int l_f(\theta,f) l_f'(\theta_0,f_0) p(Y|\theta_0,f_0,X)\, dY.$$

$l_{MP}(\theta)$ can then be approximated by

$$l_{MP}^{*}(\theta) = \frac{1}{2}\ln\left| -l_{ff}(\theta,\hat{f}_{(\theta)}) \right| - \ln\left| I_{f,f}(\theta,\hat{f}_{(\theta)};\hat{\theta},\hat{f}) \right| + l(\theta,\hat{f}_{(\theta)}). \tag{16}$$

The second approach is to find a suitable prior $p(\theta,f)$ to integrate out the nuisance parameters and obtain the posterior mode estimators, see for example Arellano and Bonhomme (2009),

$$p(\theta|Y) \propto \int_F p(\theta,f) p(Y|\theta,f,X)\, df, \tag{17}$$

where the support of f, F, is assumed to contain the true value. To find the suitable prior is very much related to the information orthogonal reparameterization method proposed by Lancaster (2002). When f is information orthogonal to θ ($I_{f\theta} = 0$), Sweeting (1995) pointed out that the log Bayesian integrated likelihood (IL) obtained from a flat prior is asymptotically equivalent to the log MPL in Cox and Reid (1987), which is (14) with $D(\theta) = 1$. In fact, this is true as long as f is weakly information orthogonal to θ as in Lemma 2.3 discussed later. If the original parameterization does not lead to information orthogonality, Lancaster (2002) suggested that one can reparameterize f as $f(g,\theta)$, where g to f is a one-one mapping, such that the new nuisance parameter g is information orthogonal to θ. To find the information orthogonal reparameterization amounts to solving the following differential equation for f,

$$\frac{\partial f}{\partial \theta} = -I_{ff}^{-1} I_{f\theta}. \tag{18}$$

The new nuisance parameter g can be recovered as the constant term in the solution. Unlike Lancaster, we will analyze the Jacobian determinant $\left|\frac{\partial g}{\partial f'}\right|$, which is a function of θ and f. Differentiating (18) with respect to g, moving $\frac{\partial f}{\partial g'}$ to the left and taking trace of both sides give

$$tr\left(\frac{\partial^2 f}{\partial\theta\partial g'}\left(\frac{\partial f}{\partial g'}\right)^{-1}\right)=\frac{\partial \ln|\frac{\partial f}{\partial g'}|}{\partial\theta}=-\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\theta}=-tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]. \tag{19}$$

where $tr(\cdot)$ is the operation to find the trace of a square matrix. The move from (18) to (19) is important since one can treat the Jacobian determinant as a prior without the need to find the information orthogonal reparameterization. The posterior or the **Jacobian integrated likelihood** (JIL) we propose can then be obtained as

$$p(\theta|Y)\propto\int_F\left|\frac{\partial g}{\partial f'}\right|p\left(Y|f,\theta,X\right)\,df. \tag{20}$$

Note that (18) can also be written as an ordinary differential equation (ODE):

$$I_{ff}^{-1}I_{f\theta}d\theta+df=0, \tag{21}$$

while (19) can be rewritten as a linear first order homogeneous partial differential equation (PDE)

$$tr\left\{\left(\frac{\partial g}{\partial f'}\right)^{-1}\left[\frac{\partial\left(\frac{\partial g}{\partial f'}\right)}{\partial f^r}I^{rs}I_{s\theta}-\frac{\partial\left(\frac{\partial g}{\partial f'}\right)}{\partial\theta}\right]\right\}=-tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]. \tag{22}$$

The matrix $\frac{\partial g}{\partial f'}$ can be interpreted as the integrating factor for (21). It may be true that even though (18) has no solutions, (19) or (22) may still have solutions as demonstrated in the example in Section 5.1. We add the assumptions below along with Assumption 1 to 4 in order to use the Jacobian prior.

Assumption 5. (i) $l(\theta,f)$ is four-times continuously differentiable with respect to the elements in f around the neighborhood of $\hat{f}_{(\theta)}$ (the MLE of f given θ), which is defined as the open ball of radius ϵ centered at $\hat{f}_{(\theta)}$ for some $\epsilon>0$, or $B_\epsilon(\hat{f}_{(\theta)})$. (ii) $|-l_{ff}(\theta,\hat{f}_{(\theta)})|>0$. (iii) Denote the support of f as F, then $|-l_{ff}(\theta,\hat{f}_{(\theta)})|^{\frac{1}{2}}\int_{F-B_\delta(\hat{f}_{(\theta)})}p(\theta,f)\frac{p(Y|\theta,f,X)}{p(Y|\theta,\hat{f}_{(\theta)},X)}df=O(T^{-1})$ with $0<\delta<\epsilon$.

The assumptions above are adapted from the analytical assumptions for the Laplace's method in Kass, Tierney, and Kadane (1990). Since we only require the relative error to be $O(T^{-1})$, our assumptions are slightly different from theirs. The following theorem shows the properties of the MPL and the JIL with appropriate priors.

Theorem 2.3. *Under Assumption 1 to 4, i) the log MPL defined in (14) satisfies:* $E\left(\frac{d\,l_{MP}(\theta)}{d\theta}\right)=O(T^{-1})$ *and* $\frac{d\ln D(\theta)}{d\theta}=tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]+O_p(T^{-\frac{1}{2}})$. *ii) The log* ***adjusted MPL*** *defined below satisfies* $E\left(\frac{\partial l_{MP}^{\dagger}(\theta)}{\partial\theta}\right)=O(T^{-1})$ *and* $\frac{d\,l_{MP}(\theta)}{d\theta}=\frac{d\,l_{MP}^{\dagger}(\theta)}{d\theta}+O_p(T^{-\frac{1}{2}})$.

$$l_{MP}^{\dagger}(\theta)=-\frac{1}{2}\ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|+\int tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]\Big|_{f=\hat{f}_{(\theta)}}d\theta+l(\theta,\hat{f}_{(\theta)}). \tag{23}$$

Under Assumption 1 to 5, iii) if $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\theta}=\frac{d\ln p(\theta,f)}{d\theta}=tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]$, *then* $E\left(\frac{d\ln p(\theta|Y)}{d\theta}\right)=O(T^{-1})$, *where $p(\theta|Y)$ is defined in (17) or (20); iv) If* $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\theta}=\frac{d\ln p(\theta,f)}{d\theta}=tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]-I^{\theta\theta}C(\theta,f)$, *where* $I^{\theta\theta}$

and $C(\theta,f)$ are defined in Theorem 2.2, then the solution for $\frac{d\ln p(\theta|Y)}{d\theta}=0$, denoted as $\tilde{\theta}$, in a neighborhood which contains the true value, satisfies $E\left(\tilde{\theta}-\theta\right)=O(T^{-2})$.

The first method eliminates the nuisance parameter by concentrating out the nuisance parameter, while the second method is through integration. The two scores $\frac{\partial l^{\dagger}_{MP}(\theta)}{\partial\theta}$ and $\frac{\partial l_{MP}(\theta)}{\partial\theta}$ are both biased up to $O(T^{-1})$, though, in general, it is easier to use the adjusted MPL, $l^{\dagger}_{MP}(\theta)$, than $l_{MP}(\theta)$, which may not have closed form, to estimate θ. The JIL score can also have $O(T^{-1})$ bias when appropriate Jacobian priors are used. If one wants to remove the $O(T^{-1})$ bias in the estimators completely, one has to add extra terms into the prior. Similarly, if one needs an MPL which can produce estimators biased up to $O(T^{-2})$, one can add $-I^{\theta\theta}C(\theta,f)$ into the integral of (23). Both the MPL and the JIL are very much related to $tr[(I_{ff}^{-1}I_{f\theta})_{/f}]$, which could be a function of θ and f. Here we define f to be **weakly information orthogonal** to θ if $tr[(I_{ff}^{-1}I_{f\theta})_{/f}]$ is at most $O(T^{-1})$. Note that f being information orthogonal to θ ($I_{f\theta}=0$) clearly implies f being weakly information orthogonal to θ but not vice versa. For example, for a linear model with an exogenous regressor, the intercept is not information orthogonal to the slope unless the sum of the related regressor's observations is 0. But the intercept is weakly information orthogonal to the slope regardless of the sum. For such a model, there is no need to solve (18) to obtain the posterior with score bias of $O(T^{-1})$ as shown in the lemma below. Note also that f being weakly information orthogonal to θ does not necessarily imply θ being weakly information orthogonal to f.

Lemma 2.3. *If f is weakly information orthogonal to θ, a prior $p(\theta,f)\propto 1$ can ensure $E\left(\frac{\partial\ln p(\theta|Y)}{\partial\theta}\right)=O(T^{-1})$ and $p(\theta|Y)\propto\exp[l^{\dagger}_{MP}(\theta)](1+O(T^{-1}))$.*

In other words, in the case of weak information orthogonality, one can use a flat prior to ensure the score to be biased up to $O(T^{-1})$ and the marginal posterior density of θ can be approximated by the exponential of the adjusted MPL.

When f is not weakly information orthogonal to θ, one has to solve (18) or (19) for $f(\theta,g)$ or $|\frac{\partial g}{\partial f'}|$. If θ is a scalar, the solution should exist in theory. The following lemma states two special cases related to the solution for $\frac{\partial g}{\partial f'}$.

Lemma 2.4. *(a) If $\frac{\partial I_{f\theta}}{\partial f'}=\frac{\partial I_{ff}}{\partial\theta}$, then $\frac{\partial g}{\partial f'}$ is equal to I_{ff} up to an arbitrary constant not involving θ and f.*
(b) If $I_{ff}^{-1}I_{f\theta}=c(\theta)+A(\theta)f$ is an affine function of f, where $A(\theta)$ ($c(\theta)$) is a matrix (vector) value function of θ, one can obtain $\left|\frac{\partial g}{\partial f'}\right|$, which is a function of only θ, by solving the following ODE,

$$\frac{d\ \ln\left|\frac{\partial g}{\partial f'}\right|}{d\theta}=tr\left[A(\theta)\right]. \tag{24}$$

Apart from the special cases, it could be difficult to find $\frac{\partial g}{\partial f'}$ in closed form. Moreover, when θ is a vector involving more than one elements, (21) may not have any solutions. For example, when the dimension of θ is 2, say, $\theta=(\theta^1,\theta^2)$, the differential of f implied by (18) may not be exact: $\frac{\partial^2 f}{\partial\theta^1\partial\theta^2}\neq\frac{\partial^2 f}{\partial\theta^2\partial\theta^1}$. There is no guarantee that (19) and (24), which become systems of PDEs when θ is a vector, will have solutions. If $\frac{\partial g}{\partial f'}$ does not exist, (23) will not be valid either, which implies neither the JIL nor the adjusted MPL exists. Arellano and Bonhomme (2009) found that a prior that reduces bias in general involves the dependent variable or the true parameter values. Extending (12) in their paper to the case when the nuisance parameter is a vector yields

$$\left|\frac{\partial g}{\partial f'}\right| \propto \left|E_{\theta_0,f_0}(l_f l_f')\right|^{-\frac{1}{2}} \left|E_{\theta_0,f_0}(-l_{ff})\right|, \tag{25}$$

where $E_{\theta_0,f_0}(\cdot)$ is the expectation taken with respect to $p(Y|\theta_0,f_0,X)$. In the light of (16), which is Severini's MPL approximation, another data dependent prior can be formulated as

$$\left|\frac{\partial g}{\partial f'}\right| \propto \left|I_{f,f}(\theta,f;\theta_0,f_0)\right|^{-1} \left|E_{\theta_0,f_0}(-l_{ff})\right|. \tag{26}$$

(25) or (26) does not satisfy (22) for arbitrary values of θ and f. The identity can only hold if θ and f are evaluated at θ_0 and f_0 on both sides. In practice, one has to drop some terms when using (25) to calculate the bias-reducing prior, see p.515 in Arellano and Bonhomme (2009). Unlike (25) and (26), which depend on the true values of the parameters, our Jacobian prior in (19) with (24) as a special case is data-independent, though it does not always exist.

If one's interest is in estimating θ as in (13), one can avoid solving the differential equations and just solve the following equation for θ,

$$dl(\theta,\hat{f}_{(\theta)}) - \frac{1}{2} d\ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right| + \sum_{\theta^r} tr\left[(I_{ff}^{-1} I_{f\theta^r})_{/f}\right]\Big|_{f=\hat{f}_{(\theta)}} d\theta^r = 0, \tag{27}$$

which is essentially the first order condition (FOC) for the JIL or the MPL up to order $O_p(1)$ when θ is a vector with r as an arbitrary index. For Case (b) in Lemma 2.4, the FOC in (27) can be modified as

$$d\ln p(Y|\theta) + \sum_{\theta^r} \frac{\partial \ln\left|\frac{\partial g}{\partial f'}\right|}{\partial \theta^r} d\theta^r = d\ln p(Y|\theta) + \sum_{\theta^r} tr\left[A_r(\theta)\right] d\theta^r = 0. \tag{28}$$

where $p(Y|\theta) = \int p\left(Y|f,\theta\right)\, df$ and $A_r(\theta) = (I_{ff}^{-1} I_{f\theta^r})_{/f}$.

3. Panel fixed effects models

In this section, we consider panel fixed effects models, and allow θ to be a vector and both $\frac{dl(\theta,\hat{f}_{(\theta)})}{d\theta}$ and $B(\theta,f)$ in (2) to be column vectors. We make the additional assumption about our data below.

Assumption 6. i) y_{it} and x_{it} are independent over i such that the log likelihood function can be written as $l(\theta,f) = \sum_{i=1}^{N} l^{(i)}(\theta,f^i) = \sum_{i=1}^{N}\sum_{t=1}^{T} \ln p(y_{it}|x_{it},\theta,f_i)$. ii) The time dimension is small relative to the cross-sectional size: $T = o(N)$.

We assume f^i is a scalar and f is an $N\times 1$ vector, though it is straightforward to allow f^i to be a vector. For all i, Assumption 1 to 5 mentioned in Section 2 now are assumed to hold for $l^{(i)}(\theta,f^i)$ with y_t and x_t replaced by y_{it} and x_{it}. For such a panel data model with fixed effects, the i-th element of f only appears in $l^{(i)}$. Hence $l_j^{(i)} = \frac{\partial l^{(i)}(\theta,f^i)}{\partial f^j} = 0$ for $i \neq j$, l_{ff} is a diagonal matrix. Sartori (2003) termed such models as models with independent stratified observations. Since the number of nuisance parameters increases with the cross-sectional sample size, Neyman and Scott (1948) and Lancaster (2000) called such parameters incidental parameters. It is well known in the literature the MLE for θ is in general inconsistent when T is fixed, which is called the incidental parameter problem. Bartolucci et al. (2016) applied the MPL in (16) to a few panel fixed effects models. Apart from the fixed effects approach, one can model the distribution of the incidental parameters (random effects models) to address the problem, see Moral-Benito (2013). One can also use suitable transformation of the dependent variable to obtain parameters with fixed dimensions transformed from the incidental parameters, see Moreira (2009).

We consider the case when N is large relative to T, which appears in many microeconomic empirical studies. This assumption is different from other econometric studies, e.g., Arellano and Bonhomme

(2009) and Hahn and Kuersteiner (2011), who assumed N and T grow to infinity at the same rate. We only consider short panel estimation rather than inference and do not assume $T \to \infty$. If $T \to \infty$, regardless of how N behaves asymptotically, the MLE for θ will be consistent and the incidental parameter problem for estimation will not exist; though, the bias due to the profile likelihood in $\sqrt{NT}(\hat{\theta}_{MLE} - \theta)$ will not disappear if $N/T > 0$ asymptotically, which will cause inference problems.

Unlike the general model in Section 2, only the bias of the profile likelihood score will affect the $O(T^{-1})$ bias of the MLE, see Arellano and Hahn (2007), since under panel fixed effects models, the second and the third term on the right hand side (RHS) in (5) will converge to 0 in probability if $T = o(N)$ (see the proof of Theorem 3.1 in Appendix A.8). Unlike Theorem 2.3, $I^{\theta\theta}C(\theta,f)$ does not affect the $O(T^{-1})$ bias of the JIL estimator asymptotically due to $plim_{N\to\infty}\frac{1}{N}I^{\theta\theta}C(\theta,f) = 0$ (converge to 0 in probability). If one uses the MPL or the JIL in the theorem below, which removes the $O(1)$ bias of the average score, one can remove the bias of order $O(T^{-1})$ in the MLE asymptotically.

Theorem 3.1. *Under Assumption 1 to 6, the asymptotic bias of the estimators from (13) and the respective first order conditions of the adjusted MPL defined in (23) and the JIL defined in (17) or (20) with the prior satisfying* $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\theta'} = \frac{d\ln p(\theta,f)}{d\theta'} = \sum_{i=1}^{N}(I^{ii}I_{i\theta})_{/f^i}$ *will converge in probability to* $O(T^{-2})$ *as* $N \to \infty$.

For panel fixed effects models, Arellano and Bonhomme (2009) showed that a flat prior can reduce bias if and only if $\frac{1}{N}\sum_{i=1}^{N}(I^{ii}I_{i\theta})_{/f^i} = o(1)$, which, albeit stricter, corroborates our Lemma 2.3. Weak information orthogonality of f to θ now requires $plim\frac{1}{N}\sum_{i=1}^{N}(I^{ii}I_{i\theta})_{/f^i}$ to be at most $O(T^{-1})$ to ensure $plim\frac{1}{N}\frac{\partial \ln p(\theta|Y)}{\partial\theta} = O(T^{-1})$ under a flat prior for θ and f.

The lemma below states the conditions under which the average profile likelihood score will only have asymptotic bias of order $O(1)$ with no lower order terms and the MPL and JIL estimators for θ are consistent.

Lemma 3.1. *Under Assumption 1 to 6, if* $l^{(i)}(\theta,f^i)$ *is a quadratic function of* f^i *and* $H_{ii} = l_{ii} + I_{ii} = 0$ *with* $l_{ii} = \frac{\partial^2 l^{(i)}(\theta,f^i)}{\partial(f^i)^2}$ *for* $i = 1,2,\ldots,N$, *then*

$$E(\hat{f}_{(\theta)} - f) = 0, \tag{29}$$

$$\underset{N\to\infty}{plim}\frac{1}{N}\frac{\partial\, l(\theta,\hat{f}_{(\theta)})}{\partial\,\theta} = \underset{N\to\infty}{plim}\frac{1}{N}B(\theta,f) = O(1), \tag{30}$$

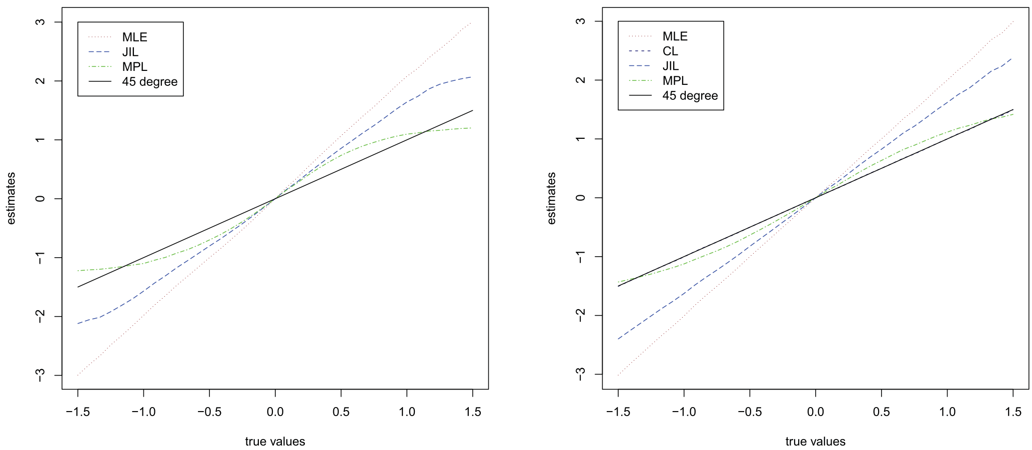
and the solution for (13) is a consistent estimator, where $B(\theta,f) = V_{\theta i,i}I^{ii} + \frac{1}{2}V_{\theta ii}I^{ii}$. *In addition, if* $B(\theta,f)$ *does not involve* f, *the estimators from the log adjusted MPL defined in (23) and the JIL defined in (20), where* $|\frac{\partial g}{\partial f'}|$ *is a function of only* θ *with* $\frac{d\,\ln|\frac{\partial g}{\partial f'}|}{d\,\theta} = I^{ii}(I_{i\theta})_{/i}$, *are the same and are consistent.*

In the next two sections, we will demonstrate our methods with two types of panel fixed effects models: static binary choice models, where the adjusted MPL and the JIL are different and their estimators are biased up to $O(T^{-2})$; and dynamic panel linear models, where the adjusted MPL estimator and the JIL estimator are the same and consistent.

3.1. Static panel binary choice models

For a panel binary choice model, the dependent variable y_{it} only takes two values: 0 or 1. Its probability of being 1 can be modeled as

$$P(y_{it} = 1|x_{it},\theta,f_i) = \Psi(f_i + x_{it}'\theta),$$



(a) probit (b) logit

Figure 1. Estimates from Different Estimators for θ when $T = 2$ and $N = 10^6$: MLE(maximum of profile likelihood), CL(conditional likelihood), JIL (Jacobian integrated likelihood) and MPL(adjusted modified profile likelihood).

where $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, x_{it} is a vector collecting all the explanatory variables and $\Psi(u)$ is the cumulative distribution function (CDF) with the probability density function (PDF) $\psi(u) = \frac{d\Psi(u)}{du}$. For a logit model, $\Psi(u)$ is the CDF of a logistic distribution with the mean equal to 0 and the variance equal to $\frac{\pi^2}{3}$. For a probit model, $\Psi(u)$ is the standard normal CDF. For both models, if x_{it} satisfies Assumption 1, 4 and 6 for all i, other assumptions required by Theorem 3.1 will also hold. If T is small, the MLE for the common parameter θ will not be consistent for large N. The bias does not only exist in $O(T^{-1})$, but also exists in higher orders such as $O(T^{-2})$ due to the nonlinear nature of the model. The log likelihood function of unit i is

$$l^{(i)} = \ln P(y_i|x_i, f_i, \theta) = \sum_{t=1}^{T} \left[y_{it} \ln \Psi(f_i + x'_{it}\theta) + (1 - y_{it}) \ln(1 - \Psi(f_i + x'_{it}\theta)) \right].$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})$ and $x_i = (x'_{i1}, x'_{i2}, \ldots, x'_{iT})'$. One can obtain

$$I_{\theta i} = \sum_t h(x'_{it}\theta + f_i)x_{it} \neq 0,$$

$$I_{ii} = \sum_t h(x'_{it}\theta + f_i),$$

where $h(u) = \frac{\psi^2(u)}{\Psi(u)[1-\Psi(u)]}$. For this model we can see that $\frac{\partial I_{\theta i}}{\partial f_i} = \frac{\partial I_{ii}}{\partial \theta}$. From Lemma 2.4, one solution for $\frac{\partial g}{\partial f'}$ can be I_{ff}, which is a diagonal matrix independent of y and satisfies the prior requirement in Theorem 3.1. Now one can use either the adjusted MPL in (23) or the Jacobian prior in (20) to obtain bias-reduced estimators for θ with asymptotic bias of order up to $O(T^{-2})$. Note that the two estimators are different for finite samples. For the static logit model, there exists a sufficient statistic for f_i: $\sum_{t=1}^{T} y_{it}$. The MLE of θ in the conditional likelihood on the sufficient statistic is consistent, see Arellano and Hahn (2007). However, such statistics do not exist for the probit model.

Figure 1 shows the estimation results for a simulation exercise where x_{it} is a scalar generated from $x_{it} = f_i + 0.3x_{i,t-1} + u_{it}$ with both f_i and u_{it} from standard normal distribution and independent of each other, and N is one million with $T = 2$. For the logit model, the line representing the MLE based on the conditional likelihood virtually overlaps the 45-degree line, which means the estimates are very close to the true values. For both models, even though the asymptotic bias of the adjusted MPL and the JIL

Table 1. Static Logit Panel.

		Bias				Standard Error			
		$N=10^2$	$N=10^3$	$N=10^4$	$N=10^5$	$N=10^2$	$N=10^3$	$N=10^4$	$N=10^5$
	t=3	0.726	0.616	0.603	0.603	0.673	0.176	0.053	0.017
MLE	t=6	0.295	0.270	0.265	0.265	0.264	0.078	0.024	0.008
	t=10	0.166	0.152	0.153	0.153	0.163	0.051	0.016	0.005
	t=3	0.312	0.255	0.247	0.248	0.463	0.126	0.038	0.012
JIL	t=6	0.071	0.057	0.053	0.053	0.204	0.062	0.019	0.006
	t=10	0.024	0.015	0.015	0.016	0.135	0.043	0.014	0.004
	t=3	0.121	0.092	0.087	0.088	0.347	0.102	0.031	0.010
MPL	t=6	0.049	0.035	0.031	0.032	0.199	0.060	0.018	0.006
	t=10	0.025	0.015	0.016	0.016	0.135	0.043	0.014	0.004
	t=3	0.054	0.006	-0.0002	0.0002	0.373	0.102	0.031	0.010
CL	t=6	0.017	0.003	-0.001	-0.0001	0.193	0.058	0.018	0.006
	t=10	0.009	-0.001	-0.0001	0.0001	0.132	0.042	0.014	0.004

MLE: maximum profile likelihood, JIL: integrated likelihood with Jacobian prior
MPL: adjusted modified profile likelihood, CL: conditional likelihood

Table 2. Static Probit Panel.

		Bias				Standard Error			
		$N=10^2$	$N=10^3$	$N=10^4$	$N=10^5$	$N=10^2$	$N=10^3$	$N=10^4$	$N=10^5$
	t=3	1.047	0.761	0.743	0.743	0.815	0.189	0.056	0.019
MLE	t=6	0.404	0.355	0.350	0.350	0.282	0.078	0.025	0.008
	t=10	0.225	0.205	0.203	0.203	0.160	0.047	0.015	0.005
	t=3	0.441	0.318	0.308	0.309	0.431	0.125	0.038	0.013
JIL	t=6	0.125	0.099	0.096	0.096	0.190	0.056	0.018	0.006
	t=10	0.059	0.046	0.044	0.044	0.126	0.037	0.012	0.004
	t=3	0.166	0.120	0.118	0.118	0.253	0.077	0.023	0.014
MPL	t=6	0.095	0.070	0.067	0.067	0.181	0.052	0.017	0.005
	t=10	0.050	0.037	0.036	0.035	0.124	0.037	0.012	0.004

MLE: maximum profile likelihood, JIL: integrated likelihood with Jacobian prior
MPL: adjusted modified profile likelihood

estimator is of order $O(T^{-2})$, these two estimators perform quite differently for $T = 2$. We can see that for the range of the true values considered, the adjusted MPL estimates are closer to the true values than those from the JIL and the line representing the MLE from the profile likelihood is the furthest away from the 45-degree line for both models. Tables 1 and 2 show the Monte Carlo results based on 1,000 simulations with the true value of θ equal to 1. We can see that the estimates under the probit model tends to have higher bias than those under the logit model. Apart from the MLE based on conditional likelihood for the logit model, the adjusted MPL appears to have the best performance in terms of both the bias and the efficiency for different sample sizes.

3.2. Dynamic panel linear models

In this section, we consider a panel linear autoregressive model (AR) with large N and small T. A panel AR(p) model can be written as

$$y_{it} = f_i + \sum_{j=1}^{p} y_{i,t-j}\rho_j + \sum_{k=1}^{K} x_{i,t,k}\beta_k + \epsilon_{it}, \tag{31}$$

We discuss two situations when $x_{it} = (x_{i,t,1}, x_{i,t,2}, \dots, x_{i,t,K})'$ is exogenous, i.e. $E(\epsilon_{it}|x_{i1}, \dots, x_{iT}) = 0$ and when $x_{i,t}$ is predetermined, i.e. ϵ_{it} can be related to the future values of the regressors. For the model to satisfy the assumptions in Theorem 3.1, x_{it} should satisfy Assumption 1, 4 and 6 for all i and k, and the roots of the characteristic polynomial should be outside the unit circle.

3.2.1. Exogenous regressors

The log likelihood of unit i conditional on the initial p periods and the exogenous regressors can be written as

$$
\begin{aligned}
l^{(i)} &= \ln p(y_i|f_i, \theta, y_{i,(1-p):0}, X_i) \\
&= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \iota f_i - Y_{i\_}\rho - X_i\beta\right)'\left(y_i - \iota f_i - Y_{i\_}\rho - X_i\beta\right)
\end{aligned} \tag{32}
$$

where $y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,T})'$, $\rho = (\rho_1, \rho_2, \dots, \rho_p)'$, $y_{i,(1-p):0}$ denotes the initial p observations of the dependent variable, ι is a vector of ones, $Y_{i\_}$ is a $T \times p$ matrix, in which the j-th row has the form $[y_{i,j-1}, y_{i,j-2}, \dots, y_{i,j-p}]$ $(j = 1, \dots, T)$ and $X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,T}]'$. One can then calculate the following expectations with respect to the conditional likelihood,

$$
\begin{aligned}
E\left(-\frac{\partial^2 l^{(i)}}{\partial f_i^2}\right) &= I_{ii} = \frac{T}{\sigma^2} = -\frac{\partial^2 l^{(i)}}{\partial f_i^2}, \\
E\left(-\frac{\partial^2 l^{(i)}}{\partial\beta\partial f_i}\right) &= I_{\beta i} = \frac{1}{\sigma^2}X_i'\iota, \\
E\left(-\frac{\partial^2 l^{(i)}}{\partial\sigma^2\partial f_i}\right) &= I_{\sigma^2 i} = E\left[\frac{\sum_{t=1}^{T}\epsilon_{it}}{\sigma^4}\right] = 0, \\
E\left(-\frac{\partial^2 l^{(i)}}{\partial\rho\partial f_i}\right) &= I_{\rho i} = \frac{1}{\sigma^2}E(Y_{i\_}'\iota), \\
&= \frac{1}{\sigma^2}\left[Th(\rho)f_i + \omega_1(X_i\beta, \rho) + \omega_2(y_{i,(1-p):0}, \rho)\right].
\end{aligned} \tag{33}
$$

$h(\rho)$ is the negative average profile likelihood score asymptotic bias with respect to ρ, whose definition along with those of $\omega_1(\cdot)$ and $\omega_2(\cdot)$, which are $p \times 1$ vector value functions not involving f_i, can be found in Appendix A.10. For this model, one can obtain (18) as

$$
\frac{\partial f_i}{\partial\beta} = -\frac{X_i'\iota}{T}, \tag{34}
$$

$$
\frac{\partial f_i}{\partial\sigma^2} = 0,
$$

$$
\frac{\partial f_i}{\partial\rho} = -h(\rho)f_i - \frac{\omega_1(X_i\beta, \rho) + \omega_2(y_{i,(1-p):0}, \rho)}{T}. \tag{35}
$$

Clearly (34) contradicts (35): $\left(\frac{\partial^2 f_i}{\partial\beta\partial\rho'}\right)' \neq \frac{\partial^2 f_i}{\partial\rho\partial\beta'}$. Though an orthogonal reparameterization does not exist for the model and (18) does not have a solution, one can see that

$$
I_{\theta i}I_{ii}^{-1} = \begin{bmatrix} I_{ii}^{-1}I_{\beta i} \\ I_{ii}^{-1}I_{\sigma^2 i} \\ I_{ii}^{-1}I_{\rho i} \end{bmatrix} = \begin{bmatrix} X_i'\iota \\ 0 \\ \frac{\omega_1(X_i\beta,\rho)+\omega_2(y_{i,(1-p):0},\rho)}{T} \end{bmatrix} + f_i \begin{bmatrix} 0 \\ 0 \\ h(\rho) \end{bmatrix}
$$

which is an affine function of f_i. To correct for the MLE bias, one can use Lemma 2.4 and obtain $\left|\frac{\partial g}{\partial f'}\right|$ by solving the following system of PDE while noting both $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\beta}$ and $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\sigma^2}$ are 0.

$$
\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\rho} = N\,h(\rho). \tag{36}
$$

$|\frac{\partial g}{\partial f'}|$ exists if (36) has a solution. The proof of the existence and the exact form of the solution can be found in Appendix A.10. Note that this Jacobian prior does not involve the dependent variable regardless of the lag order. Since $l^{(i)}$ is a quadratic function of f_i with $H_{ii}=0$, $B(\theta,f)=\frac{1}{2}(\ln I_{ii})_{/\theta}-\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\theta}$. As neither I_{ii} nor $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\theta}$ involves f, $B(\theta,f)=B(\theta)$ is a function of θ only. From Lemma 3.1, the estimators for the common parameters under the adjusted MPL and the JIL are the same and consistent. Note that the data-dependent prior estimators used by Arellano and Bonhomme (2009) have bias of order $O(T^{-2})$ for the model in this section. Our results can still be applied even if y_{it} is non-stationary, though one has to impose boundary conditions for ρ in the estimation as in Li (2015) and Dhaene and Jochmans (2016).

3.2.2. Predetermined regressors

In this section, we do not assume $x_{i,t}$ is strictly exogenous in (31). Instead we assume it is predetermined: $E(x_{i,t}\epsilon_{i,t-j})\neq 0$ for $j\geq 1$. For simplicity, we just consider $p=1$ in this subsection, though it is possible to generalize our method for $p>1$. Denote $x_i=(x'_{i,2},x'_{i,3},\ldots,x'_{i,T})'$ and z_i (with the first element equal to 1) as the collection of some observed exogenous variables, for example, $x_{i,1}$ and $y_{i,0}$, which are correlated with x_i, but uncorrelated with ϵ_i. The assumptions are summarized below.

Assumption 7. ϵ_i and x_i conditional on f_i and z_i are jointly normal with the following distributions

$$\begin{bmatrix}\epsilon_i\\ x_i\end{bmatrix}\Bigg|\, z_i,f_i\sim i.i.d.N\left(\begin{bmatrix}0\\ E(x_i|z_i,f_i)\end{bmatrix},\begin{bmatrix}\sigma^2 I_T & \Sigma_{\epsilon x}\\ \Sigma_{x\epsilon} & \Sigma_{x|z,f,}\end{bmatrix}\right), \tag{37}$$

where $\Sigma_{x|z,f,}$ and $\sigma^2I-\Sigma_{\epsilon x}\Sigma^{-1}_{x|z,f,}\Sigma_{x\epsilon}$ are positive definite, $\Sigma_{\epsilon x}=\Sigma'_{x\epsilon}$ and each row of $\Sigma_{\epsilon x}$ starts with $K\times t$ zeros for $t=0,1,\ldots T-1$.

The conditional distribution of ϵ_i on x_i and z_i is hence

$$\epsilon_i|x_i,z_i,f_i\sim N(\Sigma_{\epsilon x}\Sigma^{-1}_{x|z,f}[x_i-E(x_i|z_i,f_i)],\sigma^2I-\Sigma_{\epsilon x}\Sigma^{-1}_{x|z,f}\Sigma_{x\epsilon}). \tag{38}$$

Denote the conditional mean and variance of f_i on z_i as $\alpha' z_i$ and $\sigma^2_{f|z}$ respectively. $\Sigma_{x|z,f}$ and $E(x_i|z_i,f_i)$ can be further decomposed as

$$\Sigma_{x|z,f}=\Sigma_{x|z}-\sigma^2_{f|z}\delta\delta', \tag{39}$$

$$E\left(x_i|z_i,f_i\right)=E\left(x_i|z_i\right)+\delta f_i-\delta\alpha' z_i, \tag{40}$$

where $\delta=\frac{cov(x_i,f_i|z_i)}{\sigma^2_{f|z}}$. Due to the predetermined nature of x_i, we have $\sigma^2I-\Sigma_{\epsilon x}\Sigma^{-1}_{x|z,f}\Sigma_{x\epsilon}=\begin{bmatrix}\Omega & 0\\ 0 & \sigma^2\end{bmatrix}$ and $\Sigma_{\epsilon x}\Sigma^{-1}_{x|z,f}=\begin{bmatrix}\Gamma\\ 0\end{bmatrix}$, where Γ is a $(T-1)\times(T-1)K$ matrix and $\Omega=\sigma^2I_{T-1}-\Gamma\Sigma_{x|z,f}\Gamma'=\sigma^2I_{T-1}-\Gamma\Sigma_{x|z}\Gamma'+\sigma^2_{f|z}\zeta\zeta'$ with $\zeta=\Gamma\delta$ being a $(T-1)\times 1$ vector. We can now modify the individual log likelihood function in (32) as

$$\begin{aligned}l^{(i)}=&-\frac{T\ln 2\pi}{2}-\frac{\left(y_{i,T}-f_i-y_{i,T-1}\rho-x'_{i,T}\beta\right)^2}{2\sigma^2}-\frac{\ln\sigma^2}{2}-\frac{\ln|\Omega|}{2}\\&-\frac{1}{2}\left[y_{i,1:T-1}-(\iota-\zeta)f_i-y_{i,0:T-2}\rho-X_{i,1:T-1}\beta-\Gamma[x_i-E(x_i|z_i)]-\zeta\alpha' z_i\right]'\\&\Omega^{-1}\left[y_{i,1:T-1}-(\iota-\zeta)f_i-y_{i,0:T-2}\rho-X_{i,1:T-1}\beta-\Gamma[x_i-E(x_i|z_i)]-\zeta\alpha' z_i\right]\end{aligned} \tag{41}$$

We essentially use (37) and (38) to specify the conditional mean of ϵ_i (therefore also y_i and $\Sigma_{x|z}$, which can be estimated by linear regression. We assume such estimators are consistent. In what follows, $\tilde{x}_i$, a consistent estimator of $x_i-E(x_i|z_i)$, is used to replace $x_i-E(x_i|z_i)$ in l_i. Also note that Ω, ζ and Γ are functions of $\Sigma_{\epsilon x}$, δ and $\sigma^2_{f|z}$ and x_i instead of the conditional distribution of y_i on X_i if the SE approach were used. One can calculate the following expectations conditional on X_i, z_i and the initial observations of the dependent variable:

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial f_i^2}\right) = I_{ii} = \nu = \left[(\iota-\zeta)'\Omega^{-1}(\iota-\zeta) + \frac{1}{\sigma^2}\right],$$

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial \beta \partial f_i}\right) = I_{\beta i} = X_i'\Omega^{-1}(\iota-\zeta) + \frac{x_{i,T}}{\sigma^2},$$

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial \sigma^2 \partial f_i}\right) = I_{\sigma^2 i} = E\left[\frac{y_{i,T} - f_i - y_{i,T-1}\rho - x_{i,T}'\beta}{\sigma^4}\right] = 0,$$

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial f_i \partial vec(\Gamma)'}\right) = I_{i\Gamma} = \tilde{x}_i' \otimes [(\iota-\zeta)'\Omega^{-1}],$$

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial f_i \partial vec(\Omega)'}\right) = I_{i\Omega} = \left[E\left(y_{i,1:T-1} - (\iota-\zeta)f_i - y_{i,0:T-2}\rho - X_{i,1:T-1}\beta - \Gamma\tilde{x}_i - \zeta\alpha' z_i\right)'\Omega^{-1}\right] \otimes [(\iota-\zeta)'\Omega^{-1}] = 0,$$

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial \alpha \partial f_i}\right) = I_{\alpha i} = z_i\zeta'\Omega^{-1}(\iota-\zeta),$$

$$E\left(-\frac{\partial^2 l^{(i)}}{\partial \zeta \partial f_i}\right) = I_{\zeta i} = (f_i - \alpha' z_i)\Omega^{-1}(\iota-\zeta),$$

$$E\left(\frac{\partial^2 l^{(i)}}{\partial \rho \partial f_i}\right) = I_{\rho i} = \frac{E(y_{i,T-1})}{\sigma^2} + E[y_{i,0:T-2}'\Omega^{-1}(\iota-\zeta)].$$

Like the case with exogenous regressors, $I_{\theta i}I_{ii}^{-1}$ is also an affine function of f_i. It is easy to verify $\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\beta}$, $\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\sigma^2}$, $\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{dvec(\Gamma)'}$, $\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{dvec(\Omega)'}$ and $\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\alpha}$ are all 0. Thus one just needs to consider the following system of PDE.

$$\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\zeta} = N h_\zeta(\Sigma_{\epsilon x}, \delta, \sigma_{f|z}^2, \sigma^2) = -N\frac{\Omega^{-1}(\iota-\zeta)}{\nu}, \tag{42}$$

$$\begin{aligned}\frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\rho} &= N h_\rho(\rho, \Sigma_{\epsilon x}, \delta, \sigma_{f|z}^2, \sigma^2)\\ &= N\frac{\sum_{i=0}^{T-2}(1-\zeta_{T-i-1})\rho^i}{\nu\sigma^2} + N\sum_{i=2}^{T-1}\left(\sum_{j=i-2}^{0}(1-\zeta_{i-j-1})\rho^j\right)\frac{(\iota-\zeta)'\Omega^{-1}(,i)}{\nu}.\end{aligned} \tag{43}$$

where $\Omega^{-1}(,i)$ is the ith column of Ω^{-1} and ζ_i is the ith element in ζ. Since $\frac{d^2\ln\left|\frac{\partial g}{\partial f'}\right|}{d\zeta d\rho}$ based on (42) is not the same as $\frac{d^2\ln\left|\frac{\partial g}{\partial f'}\right|}{d\rho d\zeta}$ based on (43), there is no solution for $\ln\left|\frac{\partial g}{\partial f'}\right|$. One has to use (28) to estimate $\theta = (\alpha, \beta, \rho, \Sigma_{\epsilon x}, \delta, \sigma_{f|z}^2, \sigma^2)$, the details of which are given in Appendix A.11. Since $B(\theta, f)$ does not involve f, the estimators for the common parameters under the MPL and the JIL are again the same and consistent according to Lemma 3.1.

3.2.3. Monte Carlo evidence

In this section, we show the Monte Carlo results of four estimation methods for the linear dynamic panel: the MLE which assumes the explanatory variables are strictly exogenous, the JIL or the MPL based on

Table 3. Estimators for ρ in the Stationary Linear AR(1) Panel with Exogenous Regressors.

ρ		Bias			Standard Error		
		$N = 10^2$	$N = 10^3$	$N = 10^4$	$N = 10^2$	$N = 10^3$	$N = 10^4$
MJFOC	T=3	0.045	0.035	0.013	0.147	0.094	0.032
	T=6	0.089	0.082	0.070	0.069	0.033	0.011
SEM	T=3	0.003	0.004	0.0002	0.126	0.052	0.016
	T=6	0.006	0.001	-0.0002	0.065	0.020	0.006
MLE	T=3	-0.527	-0.524	-0.524	0.067	0.021	0.007
	T=6	-0.269	-0.268	-0.268	0.042	0.013	0.004
MJEX	T=3	-0.042	0.004	0.0005	0.118	0.059	0.017
	T=6	0.001	0.0003	-0.0002	0.065	0.020	0.006

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

the prior defined in (36) (denoted as MJEX hereafter), which also assumes the regressors are strictly exogenous, the FOC for the JIL or the MPL defined in (28) with the log Jacobian derivatives defined in (42) and (43) (MJFOC) and finally the random-effects estimator by Moral-Benito (2013), who used a simultaneous equation model (SEM) to capture the predeterminedness of the regressors. The exact details to generate x_{it}, which can be either exogenous or predetermined, are given in Appendix A.12. In all cases x_{it} is stationary and correlated with the fixed effects. We will show the results below according to whether y_{it} is stationary or not. All the results presented are based on 1,000 simulations.

3.2.3.1. Stationary case. The true values for ρ and β are, respectively, 0.5 and 0.3. y_{it} is generated from

$$y_{it} = f_i + 0.5y_{i,t-1} + 0.3x_{it} + \epsilon_{it}. \tag{44}$$

where ϵ_{it} follows a normal distribution with mean 0 and variance 4, or $\epsilon_{it} \sim i.i.d.N(0, 4)$ and $f_i \sim i.i.d.N(1, 3)$. Tables 3 and 4, respectively, show the results for ρ and β when the explanatory variable is strictly exogenous. Except the MLE for ρ and β, all other estimators appear consistent with the increase of the cross-sectional sample size (N) while the MLE bias for ρ is more pronounced than that for β. The MJFOC method, which allows the regressors to be predetermined, seems to require a very large N to reduce the bias especially for $T = 6$. The reason could be due to the increase of the common parameters with T. Note that the number of parameters in $\Sigma_{\epsilon x}$ to be estimated is $\frac{TK(T-1)}{2}$ ($K = 1$ in our experiments) and the number of parameters in δ is $(T-1)K$. When $N = 10^4$ and $T = 6$, the bias for ρ from MJFOC is still 0.07, though the bias for β is much smaller. We have increased N to one million in this case. The bias for ρ is reduced to 0.0445 with the standard error equal to 0.002. When x_{it} is predetermined as in Tables 5 and 6, the MJEX method, which assumes the regressors are exogenous, cannot always produce estimates since the FOC of the related JIL or the MPL does not have a solution for $\rho \in (-1, 1)$. The results presented are based on the simulations where MJEX has estimates. When $T = 6$ and $N = 10^4$, MJEX cannot produce estimates for all the simulations. When x_{it} is predetermined, the bias for β under MLE is more obvious and the bias for ρ under MJFOC is smaller for $N = 10^3$ and $N = 10^4$ than when x_{it} is exogenous. MJEX has smaller absolute bias for ρ than MLE, though it is more biased for β. One more thing to note is that the SEM method by Moral-Benito (2013) performs quite well in different cases in terms of both bias and efficiency.

3.2.3.2. Non-Stationary case. Though the asymptotic results in Section 3 are derived when the dependent variable is stationary, Li (2015) and Dhaene and Jochmans (2016) showed that the solution for (36) under panel fixed effects models with exogenous regressors can be applied to the case of non-stationary dependent variable. Though the asymptotic order of the bias can be different, we conjecture that our method under predetermined regressors should also work when y_{it} is non-stationary and

Table 4. Estimators for β in the Stationary Linear AR(1) Panel with Exogenous Regressors.

β		Bias			Standard Error		
		$N = 10^2$	$N = 10^3$	$N = 10^4$	$N = 10^2$	$N = 10^3$	$N = 10^4$
MJFOC	T=3	-0.024	-0.015	0.001	0.306	0.125	0.052
	T=6	0.061	-0.030	-0.004	0.158	0.075	0.034
SEM	T=3	0.032	-0.001	-0.0001	0.280	0.092	0.029
	T=6	0.015	0.001	4e-6	0.137	0.042	0.013
MLE	T=3	-0.031	-0.042	-0.04	0.142	0.046	0.014
	T=6	0.007	-0.001	-0.0001	0.088	0.030	0.009
MJEX	T=3	0.002	-0.003	-4e-5	0.155	0.050	0.016
	T=6	0.007	-0.001	0.0001	0.087	0.029	0.009

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

Table 5. Estimators for ρ in the Stationary Linear AR(1) Panel with Predetermined Regressors.

ρ		Bias			Standard Error		
		$N = 10^2$	$N = 10^3$	$N = 10^4$	$N = 10^2$	$N = 10^3$	$N = 10^4$
MJFOC	T=3	0.077	-0.011	-0.021	0.123	0.099	0.053
	T=6	0.089	0.033	0.022	0.099	0.059	0.022
SEM	T=3	-0.028	-0.004	0.001	0.119	0.056	0.025
	T=6	-0.004	0.001	0.0004	0.069	0.026	0.009
MLE	T=3	-0.539	-0.537	-0.536	0.072	0.022	0.007
	T=6	-0.266	-0.265	-0.265	0.053	0.016	0.005
MJEX	T=3	-0.005	0.115	0.190	0.120	0.065	0.019
	T=6	0.074	0.156	NA	0.070	0.044	NA

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

Table 6. Estimators for β in the Stationary Linear AR(1) Panel with Predetermined Regressors.

β		Bias			Standard Error		
		$N = 10^2$	$N = 10^3$	$N = 10^4$	$N = 10^2$	$N = 10^3$	$N = 10^4$
MJFOC	T=3	-0.084	-0.017	-0.022	0.247	0.116	0.061
	T=6	-0.093	-0.013	-0.009	0.153	0.067	0.015
SEM	T=3	-0.023	-0.006	-0.0003	0.168	0.064	0.027
	T=6	-0.008	0.001	0.0001	0.103	0.034	0.011
MLE	T=3	-0.304	-0.308	-0.307	0.112	0.035	0.011
	T=6	-0.224	-0.219	-0.220	0.071	0.022	0.007
MJEX	T=3	-0.567	-0.623	-0.659	0.131	0.051	0.013
	T=6	-0.455	-0.510	NA	0.079	0.032	NA

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

generated from the process below:

$$y_{it} = f_i + y_{i,t-1} + 0.7x_{it} + \epsilon_{it}. \tag{45}$$

The bias for ρ in general appears to be smaller for different estimators under different sample sizes than when y_{it} is stationary, as evidenced in Tables 7 and 9, where the MLE bias for ρ is much smaller than the one presented in Tables 3 and 5. The MJEX estimator for ρ seems to be fine with existent FOC solutions even when x_{it} is predetermined, which can be due to the way we generate our data. We show in Appendix A.12 that x_{it} is stationary in this case. Since y_{it} is non-stationary, the MLE bias due to the inclusion of $y_{i,t-1}$ dominates that from including x_{it}. As the Jacobian prior under MJEX is designed to remove the

Table 7. Estimators for ρ in the Non-Stationary Linear AR(1) Panel with Exogenous Regressors.

ρ		Bias			Standard Error		
		$N=10^2$	$N=10^3$	$N=10^4$	$N=10^2$	$N=10^3$	$N=10^4$
MJFOC	T=3	0.074	0.012	0.004	0.134	0.041	0.018
	T=6	0.005	0.002	0.002	0.023	0.008	0.003
SEM	T=3	0.006	-0.00001	-0.0001	0.057	0.017	0.005
	T=6	0.001	0.0001	0.0001	0.018	0.006	0.002
MLE	T=3	-0.193	-0.192	-0.192	0.045	0.014	0.004
	T=6	-0.062	-0.061	-0.061	0.017	0.005	0.002
MJEX	T=3	0.002	-0.0003	-0.0002	0.056	0.017	0.005
	T=6	0.0001	-0.00005	0.0001	0.017	0.006	0.002

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

Table 8. Estimators for β in the Non-Stationary Linear AR(1) Panel with Exogenous Regressors.

β		Bias			Standard Error		
		$N=10^2$	$N=10^3$	$N=10^4$	$N=10^2$	$N=10^3$	$N=10^4$
MJFOC	T=3	-0.024	-0.015	-0.002	0.338	0.127	0.055
	T=6	0.053	-0.032	-0.004	0.159	0.068	0.031
SEM	T=3	0.023	0.005	-0.0002	0.358	0.102	0.031
	T=6	0.013	0.001	-0.0002	0.164	0.046	0.014
MLE	T=3	0.015	0.015	0.016	0.153	0.050	0.016
	T=6	0.013	0.013	0.014	0.090	0.029	0.009
MJEX	T=3	-0.002	0.0002	0.001	0.159	0.050	0.016
	T=6	-0.0001	-0.001	0.0001	0.090	0.029	0.009

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

Table 9. Estimators for ρ in the Non-Stationary Linear AR(1) Panel with Predetermined Regressors.

ρ		Bias			Standard Error		
		$N=10^2$	$N=10^3$	$N=10^4$	$N=10^2$	$N=10^3$	$N=10^4$
MJFOC	T=3	0.028	0.012	-0.001	0.089	0.054	0.013
	T=6	0.007	-0.002	-0.0002	0.027	0.013	0.003
SEM	T=3	0.004	0.002	-0.0002	0.055	0.019	0.005
	T=6	0.002	-0.0002	-3.6e-6	0.020	0.006	0.002
MLE	T=3	-0.177	-0.174	-0.174	0.043	0.014	0.004
	T=6	-0.064	-0.063	-0.063	0.017	0.005	0.002
MJEX	T=3	0.008	0.009	0.008	0.054	0.018	0.005
	T=6	-0.006	-0.007	-0.006	0.018	0.006	0.002

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

bias due to the inclusion of $y_{i,t-1}$, it can still produce bias-reduced estimator. However, the same method will produce the biggest bias for β when $x_{i,t}$ is predetermined for different sample sizes.

4. Conclusion

We propose the Jacobian integrated likelihood (JIL) with a Jacobian prior to obtain estimators with smaller bias than that of MLE and discuss its relationship with the modified profile likelihood (MPL) in Barndorff-Nielsen and Cox (1994). We also propose the adjusted MPL, which can remove the profile

Table 10. Estimators for β in the Non-Stationary Linear AR(1) Panel with Predetermined Regressors.

β		Bias			Standard Error		
		$N=10^2$	$N=10^3$	$N=10^4$	$N=10^2$	$N=10^3$	$N=10^4$
MJFOC	T=3	-0.027	-0.018	-0.004	0.221	0.127	0.049
	T=6	-0.055	-0.010	-0.00001	0.133	0.048	0.016
SEM	T=3	0.032	0.003	0.001	0.287	0.083	0.026
	T=6	0.009	-0.001	0.0004	0.133	0.035	0.011
MLE	T=3	-0.368	-0.368	-0.366	0.132	0.042	0.013
	T=6	-0.202	-0.200	-0.199	0.081	0.023	0.008
MJEX	T=3	-0.488	-0.486	-0.483	0.140	0.043	0.014
	T=6	-0.246	-0.243	-0.241	0.082	0.023	0.008

MJFOC: FOC corrected by the Jacobian in (42) and (43)
SEM: simultaneous equation model (Moral-Benito, 2013)
MLE: maximum likelihood
MJEX: Jacobian prior based on (36) assuming the regressors are exogenous

likelihood score bias asymptotically and is easier to be computed than the original MPL. We study the incidental parameter problem in panel fixed effects models and compare the JIL and the adjusted MPL for panel probit and logit models with Monte Carlo experiments. We show how the JIL could be found when the information orthogonal reparameterization does not exist as in the linear dynamic AR(p) panel model with exogenous regressors. When the Jacobian prior cannot be found from the related partial differential equation system, neither the adjusted MPL nor the JIL exists. We demonstrate that one can still obtain consistent estimators for the common parameters by solving the first order conditions from the JIL and the adjusted MPL as for the linear AR panel model with predetermined regressors. We mainly consider the estimation issues for large N and small T. Inference studies of commonly used test statistics with incidental parameters could be the future research.

A. Appendix

A.1. Proof of Theorem 2.1

Given Assumption 1 and 2, the strong mixing sequence $\{\frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v}\}$, where $v>0$ is the element index, is a strictly stationary martingale difference sequence since $E(|\frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v}|)<\infty$ and $E(\frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v}|\mathcal{F}_{-\infty}^{t-1}) = E(\frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v}|x_t)=0$ (first order Bartlett's identity in Assumption 3). As $E[(\frac{\partial l(\eta)}{\partial \eta^v})^2] = TE[(\frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v})^2]$, where $\frac{\partial l(\eta)}{\partial \eta^v} = \sum_{t=1}^T \frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v}$, can grow to infinity with T, $\lim\limits_{T\to\infty} \frac{E[(\frac{\partial l(\eta)}{\partial \eta^v})^2]}{T} > 0$ exists. With $E[|\frac{\partial \ln p(y_t|\eta,x_t)}{\partial \eta^v}|^2]<\infty$, one can use Theorem 5.24 in White (2001) to find $\frac{\partial l(\eta)}{\partial \eta^v}/\sqrt{E[(\frac{\partial l(\eta)}{\partial \eta^v})^2]}$ converges in distribution to $N(0,1)$, and hence $\frac{\partial l(\eta)}{\partial \eta^v} = O_p(T^{\frac{1}{2}})$, or l_r and l_θ are $O_p(T^{\frac{1}{2}})$.

Expanding $\frac{\partial l(\hat\eta)}{\partial \eta^v}=0$ evaluated at the MLE estimate $(\hat\eta)$ around η gives

$$0 = \frac{\partial l(\eta)}{\partial \eta^v} + \frac{\partial^2 l(\eta)}{\partial \eta^v \partial \eta'}(\hat\eta-\eta) + \frac{1}{2}(\hat\eta-\eta)'\frac{\partial^3 l(\eta)}{\partial \eta^v \partial \eta \partial \eta'}(\hat\eta-\eta)+\dots \tag{A1}$$

One can also have $I_{\eta\eta} = E(-\frac{\partial^2 l(\eta)}{\partial \eta \partial \eta'}) = E(\frac{\partial l(\eta)}{\partial \eta}\frac{\partial l(\eta)}{\partial \eta'}) = O(T)$ (Bartlett's identity of second order) and $I_{\eta\eta}^{-1}$ exists as $E(\frac{\partial l(\eta)}{\partial \eta}\frac{\partial l(\eta)}{\partial \eta'})$ is positive definite. Since $E(|\frac{\partial^3 \ln p(y_t|\eta,x_t)}{\partial \eta^{v_1}\partial \eta^{v_2}\partial \eta^{v_3}}|)<\infty$, $\frac{\partial^3 l(\eta)}{\partial \eta^v \partial \eta \partial \eta'}$ is at most $O(T)$. Moving $\frac{\partial^2 l(\eta)}{\partial \eta^v \partial \eta'}(\hat\eta-\eta)$ to the RHS in (A1), pre-multiplying both sides by $I_{\eta\eta}^{-1}$ and performing repeated substitution on the RHS gives $\hat\eta-\eta = O_p(T^{-\frac{1}{2}}) = I_{\eta\eta}^{-1}\frac{\partial l(\eta)}{\partial \eta} + o_p(T^{-\frac{1}{2}})$.

A.2. Proof of Lemma 2.1

Given Assumption 1, $\ln p(y_t|\eta, x_t)$ is a measurable function with respect to (y_t, x_t). Due to continuous differentiability, the first five order derivatives with respect to η are measurable functions. Their σ-fields are contained in the one generated by Z_t and their mixing coefficients should be no more than $\alpha_Z(n)$, see Theorem 3.49 in White (2001). Given $\sum_{n=1}^{\infty} n[\alpha_Z(n)]^{\frac{\gamma}{9+3\gamma}} < \infty$ and $E(|\frac{\partial^4 \ln p(y_t|\eta,x_t)}{\partial\eta^{v_1}\partial\eta^{v_2}\partial\eta^{v_3}\partial\eta^{v_4}}|^{4+\gamma}) < \infty$, using Theorem 3.7 and the method to prove Lemma 10.4 in Bradley (2007), one can show that the second and the third absolute centered moments of the first four order derivatives of $l(\theta, f)$ are at most $O(T)$, e.g. $E(|H_{\theta rs}|^3) = O(T)$, where $H_{\theta rs} = l_{\theta rs} - V_{\theta rs}$ and $V_{\theta rs} = E(l_{\theta rs})$, and their fourth centered moments are at most $O(T^2)$. Due to Hölder's inequality, $E(|H_{sv}l_u|) \le \sqrt{E(H_{sv}^2)E(l_u^2)} = O(T)$ and $E(|l_r l_s l_v|) \le (E(|l_r|^3)E(|l_s|^3)E(|l_v|^3))^{\frac{1}{3}} = O(T)$. One can use Corollary 10.8 in Bradley (2007) to find that the second to the fourth order derivatives of $l(\theta, f)$ when subtracted by their respective means and divided by $\sqrt{T}$ are either normally distributed or converge to 0 asymptotically (e.g. $H_{\theta rs} = O_p(T^{\frac{1}{2}})$). Modifying Equation (5.25) in Barndorff-Nielsen and Cox (1994) gives the expansion of $\hat{f}_{(\theta)}$:

$$\hat{f}^r_{(\theta)} - f^r = I^{rs} l_s \blacktriangledown + \frac{I^{ru_1} I^{su_2} I^{tu_2} V_{u_1 u_2 u_3}}{2} l_s l_t + I^{rs} I^{tu} H_{st} l_u \blacktriangledown + O_p(T^{-\frac{3}{2}}).$$

The asymptotic expansion of the profile likelihood score around f is:

$$\frac{d\, l(\theta, \hat{f}_{(\theta)})}{d\theta} = l_\theta + l_{\theta r}(\hat{f}^r_{(\theta)} - f^r) + \frac{1}{2} l_{\theta rs}(\hat{f}^r_{(\theta)} - f^r)(\hat{f}^s_{(\theta)} - f^s) + O_p(T^{-\frac{1}{2}}) \tag{A2}$$

Substituting out $\hat{f}^r_{(\theta)} - f^r$ and replacing $l_{\theta r}$ and $l_{\theta rs}$ with $H_{\theta r} - I_{\theta r}$ and $V_{\theta rs} + H_{\theta rs}$ respectively yields (1). Taking expectation of both sides gives (3). Note that the expectations of the terms of $O_p(T^{-\frac{1}{2}})$ and $O_p(T^{-1})$ in the remainder of (1) are $O(T^{-1})$. (4) is due to the third order Bartlett's identity: $V_{\theta rs} + V_{r,\theta s} + V_{s,r\theta} = -V_{\theta,rs} - V_{r,s,\theta}$ and $V_{rst} + V_{s,rt} + V_{r,ts} = -V_{t,rs} - V_{r,s,t}$. Also note that $I^{rs}(V_{r,\theta s} + V_{s,r\theta}) = 2I^{rs}V_{r,\theta s}$ and $I^{rs}(V_{r,ts} + V_{s,rt}) = 2I^{rs}V_{r,ts}$.

A.3. Proof of Theorem 2.2

First note that

$$\left.\frac{d\, l(\theta, \hat{f}_{(\theta)})}{d\theta}\right|_{\theta=\hat{\theta}} = \frac{d\, l(\hat{\theta}, \hat{f})}{d\theta} = 0 = \frac{d\, l(\theta, \hat{f}_{(\theta)})}{d\theta} + \frac{d^2\, l(\theta, \hat{f}_{(\theta)})}{d\theta^2}(\hat{\theta} - \theta) + \frac{1}{2}\frac{d^3\, l(\theta, \hat{f}_{(\theta)})}{d\theta^3}(\hat{\theta} - \theta)^2 + O_p(T^{-\frac{1}{2}}).$$

$$\begin{aligned}
\frac{d^2\, l(\theta, \hat{f}_{(\theta)})}{d\theta^2} &= -(I_{\theta\theta} - I_{\theta r} I^{rs} I_{s\theta}) \blacktriangledown + R(\theta, f) \blacktriangledown + O_p(1). \\
\frac{d^3\, l(\theta, \hat{f}_{(\theta)})}{d\theta^3} &= W_{\theta\theta\theta} \blacktriangledown + O_p(T^{\frac{1}{2}})
\end{aligned} \tag{A3}$$

Following the arguments in the first paragraph in A.2, one can show $R(\theta, f) = O_p(T^{\frac{1}{2}})$, and then use the expansion below to find $\hat{\theta} - \theta$.

$$\begin{aligned}
\left(-\frac{d^2\, l(\theta, \hat{f}_{(\theta)})}{d\theta^2}\right)^{-1} &= (1 - I^{\theta\theta} R(\theta, f))^{-1} I^{\theta\theta} + O_p(T^{-2}) \\
&= I^{\theta\theta} + (I^{\theta\theta})^2 R(\theta, f) + (I^{\theta\theta})^3 R(\theta, f)^2 + \cdots + O_p(T^{-2}).
\end{aligned} \tag{A4}$$

$$\begin{aligned}
\hat{\theta} - \theta &= \left(-\frac{d^2\, l(\theta, \hat{f}_{(\theta)})}{d\theta^2}\right)^{-1} \frac{d\, l(\theta, \hat{f}_{(\theta)})}{d\theta} \\
&\quad + \frac{1}{2}\left(-\frac{d^2\, l(\theta, \hat{f}_{(\theta)})}{d\theta^2}\right)^{-1} \frac{d^3\, l(\theta, \hat{f}_{(\theta)})}{d\theta^3}(\hat{\theta} - \theta)^2 + O_p(T^{-\frac{3}{2}}).
\end{aligned}$$

or

$$\hat{\theta}-\theta=\left(-\frac{d^2\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^2}\right)^{-1}\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}+\frac{1}{2}\left(-\frac{d^2\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^2}\right)^{-3}\frac{d^3\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^3}\left(\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}\right)^2+O_p(T^{-\frac{3}{2}}).$$

Substituting (1), (A3) and (A4) into the above yields (5) and taking expectation of both sides give (6).

A.4. Proof of Lemma 2.2

Expanding (9) around θ gives

$$\begin{aligned}0=&\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}-\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})+\left(\frac{d^2\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^2}-\frac{d\left[\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})\right]}{d\,\theta}\right)(\tilde{\theta}-\theta)\\&+\frac{1}{2}\left(\frac{d^3\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^3}-\frac{d^2\left[\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})\right]}{d\,\theta^2}\right)(\tilde{\theta}-\theta)^2+O_p(T^{-\frac{1}{2}})\end{aligned}\tag{A5}$$

where one can substitute $\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})=\left(I^{\theta\theta}\right)^{-1}b(\theta,f)+O_p(T^{-\frac{1}{2}})$ into $\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}$ $b(\theta,\hat{f}_{(\theta)})$. Since the leading term of $\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})$ is $O(1)$, its first and second order total derivatives are also $O(1)$ and $\left(-\frac{d^2\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^2}+\frac{d\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})}{d\,\theta}\right)^{-1}$ can be expanded as the RHS of (A4) with different $O_p(T^{-2})$ terms. Hence (A5) can be rewritten as

$$\begin{aligned}\tilde{\theta}-\theta=&\left(-\frac{d^2\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^2}+\frac{d\left[\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})\right]}{d\,\theta}\right)^{-1}\left[\left(\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}-\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})\right)\right.\\&\left.+\frac{1}{2}\left(\frac{d^3\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^3}-\frac{d^2\left[\left(I^{\theta\theta}(\theta,\hat{f}_{(\theta)})\right)^{-1}b(\theta,\hat{f}_{(\theta)})\right]}{d\,\theta^2}\right)(\tilde{\theta}-\theta)^2\right]+O_p(T^{-\frac{3}{2}})\\=&I^{\theta\theta}\left(\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}-\left(I^{\theta\theta}\right)^{-1}b(\theta,f)\right)+(I^{\theta\theta})^2R(\theta,f)\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}\\&+\frac{1}{2}I^{\theta\theta}\frac{d^3\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^3}(\tilde{\theta}-\theta)^2+O_p(T^{-\frac{3}{2}})\\=&I^{\theta\theta}\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}-b(\theta,f)+(I^{\theta\theta})^2R(\theta,f)\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}\\&+\frac{1}{2}\left(I^{\theta\theta}\right)^3\frac{d^3\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta^3}\left(\frac{d\, l(\theta,\hat{f}_{(\theta)})}{d\,\theta}\right)^2+O_p(T^{-\frac{3}{2}})\end{aligned}\tag{A6}$$

Substituting (1) and (A3) into the above yields (10). Taking expectations of both sides gives (11).

A.5. Proof of Theorem 2.3

Note that

$$
\begin{aligned}
\frac{d\ \ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta} &= tr\left[-l_{ff}^{-1}(\theta,\hat{f}_{(\theta)})\left(-l_{ff\theta}(\theta,\hat{f}_{(\theta)})-l_{fff}(\theta,\hat{f}_{(\theta)})\frac{\partial\hat{f}_{(\theta)}}{\partial\theta}\right)\right],\\
&= tr\left[-\left(I_{ff}+O_p(T^{\frac{1}{2}})\right)^{-1}\left(V_{ff\theta}-V_{fff}l_{ff}^{-1}(\theta,\hat{f}_{(\theta)})l_{f\theta}(\theta,\hat{f}_{(\theta)})+O_p(T^{\frac{1}{2}})\right)\right],\\
&= tr\left[-\left(I_{ff}^{-1}+O_p(T^{-\frac{3}{2}})\right)\left(V_{ff\theta}-V_{fff}I_{ff}^{-1}I_{f\theta}+O_p(T^{\frac{1}{2}})\right)\right],\\
&= tr\left[-I_{ff}^{-1}\left(V_{ff\theta}-V_{fff}I_{ff}^{-1}I_{f\theta}\right)+O_p(T^{-\frac{1}{2}})\right].
\end{aligned}
\tag{A7}
$$

For the mixed log model derivatives, define $\hat{l}_{R_1;R_2}=l_{R_1;R_2}(\hat{\theta},\hat{f};\hat{\theta},\hat{f})$, where R_1 and R_2 are arbitrary index sets. From (5.75) and (5.83) in Barndorff-Nielsen and Cox (1994), note that

$$
\hat{l}_{R_1;R_2}=l_{R_1;R_2}(\theta,f;\theta,f)+O_p(T^{\frac{1}{2}})=V_{R_1;R_2}+O_p(T^{\frac{1}{2}})=\sum_{k=1}^{|R_2|}\sum_{R_2/k}V_{R_1,R_{21},\ldots,R_{2k}}+O_p(T^{\frac{1}{2}}).
$$

Also note that

$$
l_{R_1;R_2}(\theta,\hat{f}_{(\theta)};\hat{\theta},\hat{f})=\hat{l}_{R_1;R_2}+O_p(T^{\frac{1}{2}})=V_{R_1;R_2}+O_p(T^{\frac{1}{2}}).
$$

Using the above, one can have

$$
\begin{aligned}
\frac{d\ \ln\left|l_{f;\hat{f}}(\theta,\hat{f}_{(\theta)})\right|}{d\theta} &= tr\left[l_{f;\hat{f}}^{-1}(\theta,\hat{f}_{(\theta)})\left(l_{f\theta;\hat{f}}(\theta,\hat{f}_{(\theta)})+l_{ff;\hat{f}}(\theta,\hat{f}_{(\theta)})\frac{\partial\hat{f}_{(\theta)}}{\partial\theta}\right)\right]\\
&= tr\left[\left(V_{f;f}+O_p(T^{\frac{1}{2}})\right)^{-1}\left(V_{f\theta;f}-V_{ff;f}I_{ff}^{-1}I_{f\theta}+O_p(T^{\frac{1}{2}})\right)\right],\\
&= tr\left[I_{ff}^{-1}\left(V_{f\theta;f}-V_{ff;f}I_{ff}^{-1}I_{f\theta}\right)+O_p(T^{-\frac{1}{2}})\right];
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{d\ \ln D(\theta)}{d\theta} &= \frac{d\ \ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta}-\frac{d\ \ln\left|l_{f;\hat{f}}(\theta,\hat{f}_{(\theta)})\right|}{d\theta},\\
&= tr\left[-I_{ff}^{-1}\left(V_{ff\theta}+V_{f\theta;f}-(V_{fff}+V_{ff;f})I_{ff}^{-1}I_{f\theta}\right)+O_p(T^{-\frac{1}{2}})\right],\\
&= tr\left[I_{ff}^{-1}\left((I_{f\theta})_{/f}-(I_{ff})_{/f}I_{ff}^{-1}I_{f\theta}\right)+O_p(T^{-\frac{1}{2}})\right],\\
&= tr\left[\left(I_{ff}^{-1}I_{f\theta}\right)_{/f}\right]+O_p(T^{-\frac{1}{2}}).
\end{aligned}
$$

The total derivative of the log modified profile likelihood can be expanded as

$$
\begin{aligned}
\frac{d\,l_{MP}(\theta)}{d\theta} &= -\frac{1}{2}\frac{d\ \ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta}+\frac{d\ \ln D(\theta)}{d\theta}+\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}\\
&= \frac{d\,l_{MP}^{\dagger}(\theta)}{d\theta}+O_p(T^{-\frac{1}{2}}),\\
&= \frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}-B(\theta,f)+O_p(T^{-\frac{1}{2}}),
\end{aligned}
\tag{A8}
$$

Taking expectation of both sides, one can see that $E\left(\frac{d\,l_{MP}(\theta)}{d\theta}\right) = O(T^{-1})$. For (17), one can use the Laplace's method to expand it as below,

$$\begin{aligned} p(\theta|Y) &\propto \int_F p(\theta,f)\exp\left[T\frac{l(\theta,f)}{T}\right]df, \\ &\propto p(\theta,\hat{f}_{(\theta)})\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|^{-\frac{1}{2}} p(Y|\theta,\hat{f}_{(\theta)})\left(1+O(T^{-1})\right), \end{aligned} \tag{A9}$$

Taking log and derivative of both sides gives

$$\begin{aligned} \frac{d\,\ln p(\theta|Y)}{d\theta} &= \frac{d\,\ln p(\theta,\hat{f}_{(\theta)})}{d\theta} - \frac{1}{2}\frac{d\,\ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta} + \frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta} + O(T^{-1}), \\ &= \frac{d\,\ln p(\theta,f)}{d\theta} - \frac{1}{2}\frac{d\,\ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta} + \frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta} + O_p(T^{-\frac{1}{2}}). \end{aligned} \tag{A10}$$

Hence from (A7) and (2) if $\frac{d\ln|\frac{\partial g}{\partial f'}|}{d\theta} = \frac{d\ln p(\theta,f)}{d\theta} = tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]$, then $E\left(\frac{d\,\ln p(\theta|Y)}{d\theta}\right) = O(T^{-1})$. Expanding $\frac{d\,\ln p(\tilde{\theta}|Y)}{d\theta} = 0$ around θ yields

$$0 = \frac{d\,\ln p(\theta|Y)}{d\theta} + \frac{d^2\,\ln p(\theta|Y)}{d\theta^2}(\tilde{\theta}-\theta) + \frac{1}{2}\frac{d^3\,\ln p(\theta|Y)}{d\theta^3}(\tilde{\theta}-\theta)^2 + O_p(T^{-\frac{1}{2}})$$

or

$$(\tilde{\theta}-\theta) = \left(-\frac{d^2\,\ln p(\theta|Y)}{d\theta^2}\right)^{-1}\left[\frac{d\,\ln p(\theta|Y)}{d\theta} + \frac{1}{2}\frac{d^3\,\ln p(\theta|Y)}{d\theta^3}(\tilde{\theta}-\theta)^2\right] + O_p(T^{-\frac{3}{2}}) \tag{A11}$$

If $\frac{d\ln p(\theta,f)}{d\theta} = tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right] - I^{\theta\theta}C(\theta,f)$, the leading terms of both $\frac{d\ln p(\theta,f)}{d\theta}$ and $\frac{d\,\ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta}$ are $O(1)$. Hence $\frac{d^2\,l(\theta,\hat{f}_{(\theta)})}{d\theta^2}$ dominates $\frac{d^2\,\ln p(\theta|Y)}{d\theta^2}$ asymptotically. Using (A4), (A7) and (A10), one can rewrite the RHS of (A11) as in the RHS of the last equal sign in (A6) albeit with different $O_p(T^{-\frac{3}{2}})$ terms. Incorporating (1) and (A3) and taking expectations gives $E(\tilde{\theta}-\theta) = O(T^{-2})$.

A.6. Proof of Lemma 2.3

If f is weakly information orthogonal to θ: i.e. $tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right]$ is at most $O(T^{-1})$, from Theorem 2.3, one can choose the prior to be $\ln\left|\frac{\partial g}{\partial f'}\right| = 0$ or $\left|\frac{\partial g}{\partial f'}\right| = 1$ to ensure $E\left(\frac{d\,\ln p(\theta|Y)}{d\theta}\right) = O(T^{-1})$. Substituting the prior into (A9) and using the Laplace's method yields

$$p(\theta|Y) \propto \left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|^{-\frac{1}{2}} p(Y|\theta,\hat{f}_{(\theta)})\left(1+O(T^{-1})\right).$$

Taking exponential of (23) with $tr\left[(I_{ff}^{-1}I_{f\theta})_{/f}\right] = O(T^{-1})$ gives the result.

A.7. Proof of Lemma 2.4

For Case (a), if $\frac{\partial I_{f\theta}}{\partial f'} = \frac{\partial I_{ff}}{\partial\theta}$, then $I_{f\theta}d\theta + I_{ff}df = 0$ is exact and $\frac{\partial g}{\partial f'} = I_{ff}$ is therefore an integrating factor for (21).

For Case(b), if $I_{ff}^{-1}I_{f\theta} = c(\theta) + A(\theta)f$, from (19) one can see that $tr(\frac{\partial I_{ff}^{-1}I_{f\theta}}{\partial f'}) = tr[A(\theta)] = \frac{d\,\ln|\frac{\partial g}{\partial f'}|}{d\theta}$. One possible solution for $|\frac{\partial g}{\partial f'}|$ is a function of only θ but not f, which can be obtained by solving the ODE in (24).

A.8. Proof of Theorem 3.1

The asymptotic expansions shown in Section 2 still hold here with all the super/sub-scripts equal to i and different asymptotic orders due to the incidental parameters. Take (1) as an example. Note that for all i, $E(l_i^2) = O(T)$, $E(l_i^4) = O(T^2)$ (Appedix A.2 and $l_i = \frac{\partial \sum_{t=1}^T \ln p(y_{it}|x_{it},\theta,f^i)}{\partial f^i}$) and $E[(\frac{1}{N}\sum_{i=1}^N l_i)^2] = \frac{1}{N^2}\sum_{i=1}^N E(l_i^2) = O(TN^{-1})$ given Assumption 6. If $T = o(N)$, $\lim_{N\to\infty} E[(\frac{1}{N}\sum_{i=1}^N l_i)^2] = 0$ and hence $\frac{1}{N}\sum_{i=1}^N l_i \xrightarrow{p} 0$ and $\frac{1}{N} I_{\theta i} I^{ii} l_i \xrightarrow{p} 0$. Similarly, $\frac{1}{N} l_\theta$, $\frac{1}{N}(l_{\theta\theta} - V_{\theta\theta})$, $\frac{1}{N}(l_{\theta\theta\theta} - V_{\theta\theta\theta})$ and $\frac{1}{N} R(\theta,f)$ defined in (8) all converge in probability to 0. Since $E(l_i^2 H_{i\theta}^2) \le \sqrt{E(l_i^4)E(H_{i\theta}^4)} = O(T^2)$, the summands in $l_i I^{ii} H_{i\theta}$ have finite second moments. One can then use Corollary 3.9 in White (2001) to show $\frac{1}{N}(l_i I^{ii} H_{i\theta} - V_{\theta i,i} I^{ii}) \xrightarrow{p} 0$. Similarly, the terms on the RHS between the first two ▼ in (1) are $O(N)$. For the remaining terms, note that $E[(l_i^4)^{1+\frac{\gamma}{4}}] = E(|l_i|^{4+\gamma}) \le [\sum_{t=1}^T E^{\frac{1}{4+\gamma}}(|\frac{\partial \ln p(y_{it}|x_{it},\theta,f_i)}{\partial f_i}|^{4+\gamma})]^{4+\gamma} < \infty$ for finite T by Minkowski's inequality and Assumption 2, and one also can apply law of large numbers to terms analogical to $T^r l_i^4$ with $r \le -2$. One can then have

$$
\underset{N\to\infty}{plim}\frac{1}{N}\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\,\theta'} = \underset{N\to\infty}{plim}\frac{B(\theta,f)'}{N}\,\blacktriangledown + O\left(T^{-1}\right) \tag{A12}
$$

$$
\begin{aligned}
B(\theta,f) &= I^{ii}V_{i,i\theta} - \frac{1}{2}V_{iii}\left(I^{ii}\right)^2 I_{i\theta} - V_{ii,i}\left(I^{ii}\right)^2 I_{i\theta} + \frac{1}{2}I^{ii}V_{ii\theta} \\
&= I^{ii}\left[V_{i,i\theta} + V_{ii\theta} - (V_{iii} + V_{ii,i})I^{ii}I_{i\theta} - \frac{1}{2}\left(V_{ii\theta} - V_{iii}I^{ii}I_{i\theta}\right)\right] \\
&= -\sum_{i=1}^{N}(I^{ii}I_{i\theta})_{f^i} - \frac{1}{2}I^{ii}(V_{ii\theta} - V_{iii}I^{ii}I_{i\theta}).
\end{aligned}
$$

Denote the solution for (13) as $\tilde{\theta}$. One can have

$$
\begin{aligned}
0 &= \left.\left(\frac{d\,l(\theta,\hat{f}_\theta)}{d\,\theta^r} - B^r(\theta,\hat{f}_{(\theta)})\right)\right|_{\theta=\tilde{\theta}} \\
&= \frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\,\theta^r} - B^r(\theta,\hat{f}_{(\theta)}) + \left[\frac{d^2\,l(\theta,\hat{f}_{(\theta)})}{d\theta^r\,d\theta'} - \frac{dB^r(\theta,\hat{f}_{(\theta)})}{d\theta'}\right](\tilde{\theta}-\theta) \\
&\quad + \frac{1}{2}(\tilde{\theta}-\theta)'\left[\frac{d^3\,l(\theta,\hat{f}_{(\theta)})}{d\theta^r\,d\theta\,d\theta'} - \frac{d^2B^r(\theta,\hat{f}_{(\theta)})}{d\theta\,d\theta'}\right](\tilde{\theta}-\theta) + \ldots
\end{aligned} \tag{A13}
$$

where θ^r and $B^r(\theta,f)$ are the rth element of θ and $B(\theta,f)$ respectively. Note:

$$
\underset{N\to\infty}{plim}\frac{1}{N}\left[\frac{dl(\theta,\hat{f}_{(\theta)})}{d\,\theta} - B(\theta,\hat{f}_{(\theta)}) - \left(\frac{dl(\theta,\hat{f}_{(\theta)})}{d\,\theta} - B(\theta,f)\right)\right] = O(T^{-1}) \tag{A14}
$$

$$
\left[\underset{N\to\infty}{plim}\frac{1}{N}\left(\frac{dB(\theta,\hat{f}_{(\theta)})}{d\theta'} - \frac{d^2\,l(\theta,\hat{f}_{(\theta)})}{d\theta\,d\theta'}\right)\right]^{-1} = \underset{N\to\infty}{plim}\,NI^{\theta\theta} + O(T^{-2}), \tag{A15}
$$

where $I^{\theta\theta} = (I_{\theta\theta} - I_{\theta i}I^{ii}I_{i\theta})^{-1} = O(N^{-1}T^{-1})$.From (A12) and (A13),we can obtain

$$
\underset{N\to\infty}{plim}\,(\tilde{\theta}^r - \theta^r) = \underset{N\to\infty}{plim}\,I^{\theta^r\theta}\left(\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\,\theta} - B(\theta,f)\right) + O(T^{-2}) = O(T^{-2}).
$$

For panel fixed effects models, (A8) can be rewritten as

$$\begin{aligned}\underset{N\to\infty}{plim}\frac{1}{N}\frac{d\,l_{MP}(\theta)}{d\theta} &= \underset{N\to\infty}{plim}\frac{1}{N}\frac{d\,l^{\dagger}_{MP}(\theta)}{d\theta} + O(T^{-1}),\\ &= \underset{N\to\infty}{plim}\frac{1}{N}\left(\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta} - B(\theta,f)\right) + O(T^{-1}) = O(T^{-1}).\end{aligned}$$

Hence $plim\frac{1}{N}\frac{d\,l_{MP}(\theta)}{d\theta}$ and $plim\frac{1}{N}\frac{d\,l^{\dagger}_{MP}(\theta)}{d\theta}$ are of order $O(T^{-1})$. (A10) can be rewritten as

$$\underset{N\to\infty}{plim}\frac{1}{N}\frac{d\,\ln p(\theta|Y)}{d\theta} = \underset{N\to\infty}{plim}\frac{1}{N}\left[\frac{d\,\ln p(\theta,f)}{d\theta} - \frac{1}{2}\frac{d\,\ln\left|-l_{ff}(\theta,\hat{f}_{(\theta)})\right|}{d\theta} + \frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta}\right] + O(T^{-1}).$$

Therefore, if $plim\frac{1}{N}[\frac{d\ln p(\theta,f)}{d\theta'} - \sum_{i=1}^{N}(I^{ii}I_{i\theta})_{/f^i}] = O(T^{-1})$, then $plim\frac{1}{N}\frac{d\,\ln p(\theta|Y)}{d\theta}$ will also be $O(T^{-1})$. The bias of the estimators from the respective FOCs of $l_{MP}(\theta)$, $l^{\dagger}_{MP}(\theta)$ and $\ln p(\theta|Y)$, whose average scores are all free of the $O(1)$ bias, will therefore converge in probability to $O(T^{-2})$.

A.9. Proof of Lemma 3.1

Since $l^{(i)}(\theta,f^i)$ is a quadratic function of f^i, its derivative with respect to f^i of order higher than 2 is 0. One can have $l^{(i)}_i(\theta,\widehat{f^i_{(\theta)}}) = 0 = l_i(\theta,f^i) + l_{ii}(\theta,f^i)(\hat{f}^i_{(\theta)} - f^i)$ and $(\hat{f}^i_{(\theta)} - f^i) = I^{ii}l_i$ with $I^{ii} = (-l_{ii})^{-1}$. Taking expectation gives (29). One can also have $B(\theta,\hat{f}_{(\theta)}) = B(\theta,f) + B_iI^{ii}l_i$, where $B_i = \frac{\partial B(\theta,f)}{\partial f^i} = 0$ when $B(\theta,f)$ does not involve f, and $l(\theta,\hat{f}_\theta) = l(\theta,f) + \frac{1}{2}I^{ii}l_il_i$. Note that the terms of $O(T^{-1})$ in (A14) does not exist here and

$$\begin{aligned}\underset{N\to\infty}{plim}\frac{1}{N}\frac{dl(\theta,\hat{f}_\theta)}{d\theta} &= \underset{N\to\infty}{plim}\frac{1}{N}\left[l_\theta(\theta,f) + l_{\theta i}l_iI^{ii} - \frac{1}{2}I^{ii}(-l_{\theta ii})I^{ii}l_il_i\right]\\ &= \underset{N\to\infty}{plim}\frac{1}{N}\left[V_{\theta i,i}I^{ii} + \frac{1}{2}I^{ii}V_{ii\theta}\right]\\ &= \underset{N\to\infty}{plim}\frac{1}{N}B(\theta,f) = \underset{N\to\infty}{plim}\frac{1}{N}B(\theta,\hat{f}_{(\theta)})\end{aligned}$$

where $V_{ii\theta} = E(l_{ii\theta}) = l_{ii\theta}$ if $l_{ii} = E(l_{ii})$. Using (A13) and (A15) gives $\underset{N\to\infty}{plim}(\tilde{\theta} - \theta) = \underset{N\to\infty}{plim}\, I^{\theta\theta}[\frac{d\,l(\theta,\hat{f}_{(\theta)})}{d\theta} - B(\theta,\hat{f}_{(\theta)})] = 0$. The solution for (13) is consistent for θ.

If $B(\theta,f) = B(\theta)$ does not involve f, which implies $V_{\theta i,i}$ not involving f when $l^{(i)}(\theta,f^i)$ is a quadratic function of f^i, $tr[(I^{-1}_{ff}I_{f\theta})_{/f}] = I^{ii}(I_{i\theta})_{/i} = I^{ii}(-V_{i\theta i} - V_{i\theta,i}) = \frac{d\,\ln|\frac{\partial g}{\partial f'}|}{d\theta}$ will not involve f and hence $|\frac{\partial g}{\partial f'}|$, a function of θ only, can be taken outside the integral in (20). Since the likelihood function is quadratic in f, using the Laplace method to integrate out f in (20) will produce the exact result and taking log will give the log MPL in (23). Differentiating the log JIL, $\ln p(\theta|Y)$, or the log adjusted MPL, $l^{\dagger}_{MP}(\theta)$, yields $\frac{dl(\theta,\hat{f}_{(\theta)})}{d\theta} - B(\theta)$. Therefore the estimators from the JIL and the adjusted MPL are the same as the solution for (13), which is consistent.

A.10. Solution for (36)

Through repeated substitution, one can rewrite the model in (31) as

$$[\mathbf{y}'_{i,-p}, y_{i,1}, y_{i,2},\ldots,y_{i,T-1}]' = f_ic_1 + I_{T-1+p}\otimes y'_{i,-p}c_2 + CX_i\beta + C\epsilon_i$$

$$\underset{p\times 1}{\mathbf{y}_{i,-p}} = \begin{pmatrix} y_{i,-p+1} \\ y_{i,-p+2} \\ \dots \\ y_{i,-1} \\ y_{i,0} \end{pmatrix}, \underset{p\times p}{P} = \begin{pmatrix} \rho_1 & 1 & 0 & \dots 0 \\ \rho_2 & 0 & 1 & \dots 0 \\ \dots & \dots\dots & \dots\dots & \\ \rho_{p-1} & 0 & 0 & \dots 1 \\ \rho_p & 0 & 0 & \dots 0 \end{pmatrix},$$

$$\underset{(T-1+p)\times 1}{c_1} = \begin{pmatrix} 0_{p\times 1} \\ 1 \\ P_{(1,1)}+1 \\ P^2_{(1,1)}+P_{(1,1)}+1 \\ \dots \\ P^{T-2}_{(1,1)}+P^{T-3}_{(1,1)}+\dots+P_{(1,1)}+1 \end{pmatrix}, \underset{[p^2+(T-1)p]\times 1}{c_2} = \begin{pmatrix} vec(I_p) \\ P_{(,1)} \\ P^2_{(,1)} \\ \dots \\ P^{T-1}_{(,1)} \end{pmatrix},$$

$$\underset{(T-1+p)\times T}{C} = \begin{pmatrix} 0_{p\times 1} & 0_{p\times 1} & \dots 0_{p\times 1} & 0_{p\times 1} \\ 1 & 0 & \dots 0 & 0 \\ P_{(1,1)} & 1 & \dots 0 & 0 \\ \dots & \dots\dots & \dots\dots & \\ P^{T-2}_{(1,1)} & P^{T-3}_{(1,1)} & \dots 1 & 0 \end{pmatrix}. \tag{A16}$$

where $P^n_{(1,1)}$ and $P^n_{(,1)}$ denote the (1,1) element and the first column of the matrix P to power n. To find $E(Y'_{i\_}\iota)$, one can make use of (A16). We define $h : R^p \mapsto R^p$, $\omega_1 : R^{p+T} \mapsto R^p$ and $\omega_2 : R^{2p} \mapsto R^p$ respectively as

$$\underset{p\times 1}{h}(\underset{p\times 1}{\rho}) = \frac{1}{T}\begin{pmatrix} \iota' c_{1(p:T+p-1)} \\ \iota' c_{1(p-1:T+p-2)} \\ \dots \\ \iota' c_{1(1:T)} \end{pmatrix} = \frac{1}{T}\begin{pmatrix} \iota' C_{(p:T+p-1,)}\iota \\ \iota' C_{(p-1:T+p-2,)}\iota \\ \dots \\ \iota' C_{(1:T,)}\iota \end{pmatrix},$$

$$\underset{p\times 1}{\omega_1}(\underset{T\times 1}{X_i\beta}, \underset{p\times 1}{\rho}) = \begin{pmatrix} \iota' C_{(p:T+p-1)}X_i\beta \\ \iota' C_{(p-1:T+p-2)}X_i\beta \\ \dots \\ \iota' C_{(1:T)}X_i\beta \end{pmatrix},$$

$$\underset{p\times 1}{\omega_2}(\underset{p\times 1}{y_{i,-p}}, \underset{p\times 1}{\rho}) = \begin{pmatrix} \iota'(I_{T-1+p}\otimes y'_{i,-p}c_2)_{p:T+p-1} \\ \iota'(I_{T-1+p}\otimes y'_{i,-p}c_2)_{p-1:T+p-2} \\ \dots \\ \iota'(I_{T-1+p}\otimes y'_{i,-p}c_2)_{1:T} \end{pmatrix},$$

where $a_{1:T}$ and $A_{(1:T,)}$ denote the 1 to T elements and the 1 to T rows of a and A respectively. Note that since $E(C\epsilon_i)$ is equal to zero, we can obtain $E(Y'_{i\_}\iota) = \left[Th(\rho)f_i + \omega_1(X_i\beta, \rho) + \omega_2(y_{i,-p}, \rho)\right]$ and hence (33). Also note that the rth element in $h(\rho)$ can be written as

$$h_r(\rho) = \frac{T-r}{T} + \frac{T-r-1}{T}P_{(1,1)} + \dots + \frac{1}{T}P^{T-r-1}_{(1,1)} \qquad for \quad r = 1, \dots, p. \tag{A17}$$

Equation (36) implies,

$$\frac{1}{N} d\ \ln\left|\frac{\partial g}{\partial f'}\right| = \sum_{k=1}^{p} h_k(\rho)\, d\rho_k, \tag{A18}$$

where $\left|\frac{\partial g}{\partial f'}\right|$ is a function of ρ only. To prove that $\ln\left|\frac{\partial g}{\partial f'}\right|$ exists, we can prove its differential is exact. Before the proof, we establish the lemma below.

Lemma A.1.

$$\frac{\partial P^{r+j}_{(1,1)}}{\partial \rho_r} = \frac{\partial P^{s+j}_{(1,1)}}{\partial \rho_s} \tag{A19}$$

where $r, s = 1, 2, \dots, p$ and j can be zero or any positive integer.

Proof. It is obvious that if $r = s$, Eq. (A19) holds. Without loss of generality, we can assume $r < s$. Define $P^{n-k}_{(1,1)} = 1$ if $n - k = 0$ and $P^{n-k}_{(1,1)} = 0$ if $n - k < 0$. One can have

$$P^{n}_{(1,1)} = \sum_{k=1}^{p} \rho_k P^{n-k}_{(1,1)}.$$

The above equation implies $\frac{\partial P^{n}_{(1,1)}}{\partial \rho_r} = 0$ and $\frac{\partial P^{n}_{(1,1)}}{\partial \rho_r} = 1$ for $n < r$ and $n = r$, respectively. Then we can prove (A19) by mathematical induction, which involves the following three steps:

1. We assume that $\frac{\partial P^{r+j-k}_{(1,1)}}{\partial \rho_r} = \frac{\partial P^{s+j-k}_{(1,1)}}{\partial \rho_s}$ holds for any positive integer k. The left and right hand side of (A19) can be rewritten as

$$\frac{\partial P^{r+j}_{(1,1)}}{\partial \rho_r} = \rho_1 \frac{\partial P^{r+j-1}_{(1,1)}}{\partial \rho_r} + \cdots + \frac{\partial \left(\rho_r P^{r+j-r}_{(1,1)} \right)}{\partial \rho_r} + \cdots + \rho_s \frac{\partial P^{r+j-s}_{(1,1)}}{\partial \rho_r} + \cdots + \rho_p \frac{\partial P^{r+j-p}_{(1,1)}}{\partial \rho_r},$$
$$\frac{\partial P^{s+j}_{(1,1)}}{\partial \rho_s} = \rho_1 \frac{\partial P^{s+j-1}_{(1,1)}}{\partial \rho_s} + \cdots + \rho_r \frac{\partial P^{s+j-r}_{(1,1)}}{\partial \rho_s} + \cdots + \frac{\partial \left(\rho_s P^{s+j-s}_{(1,1)} \right)}{\partial \rho_s} + \cdots + \rho_p \frac{\partial P^{s+j-p}_{(1,1)}}{\partial \rho_s}.$$

   Given our assumption, if the above two are the same, the following must hold,

$$P^{j}_{(1,1)} + \rho_r \frac{\partial P^{r+j-r}_{(1,1)}}{\partial \rho_r} + \rho_s \frac{\partial P^{r+j-s}_{(1,1)}}{\partial \rho_r} = P^{j}_{(1,1)} + \rho_r \frac{\partial P^{s+j-r}_{(1,1)}}{\partial \rho_s} + \rho_s \frac{\partial P^{s+j-s}_{(1,1)}}{\partial \rho_s},$$

   which is obviously true. Hence if $\frac{\partial P^{r+j-k}_{(1,1)}}{\partial \rho_r} = \frac{\partial P^{s+j-k}_{(1,1)}}{\partial \rho_s}$ holds, then $\frac{\partial P^{r+j}_{(1,1)}}{\partial \rho_r} = \frac{\partial P^{s+j}_{(1,1)}}{\partial \rho_s}$ is also true.
2. The smallest possible number for j is 0, which indicates both sides of (A19) are equal to each other and equal to 1.
3. From the above two points, we know that Lemma A.1 is true in general.

□

Now we are ready to prove that there exists a solution for the partial differential equation system in (A18).

Proof. To prove the differential of $\ln \left| \frac{\partial g}{\partial f'} \right|$ to be exact, one can prove

$$\frac{\partial h_r(\rho)}{\partial \rho_s} = \frac{\partial h_s(\rho)}{\partial \rho_r}, \tag{A20}$$

where $h_r(\rho)$ is defined in (A17). For (A20) to hold, one should have

$$\frac{T-r-s}{T} \frac{\partial P^{s}_{(1,1)}}{\partial \rho_s} + \cdots + \frac{1}{T} \frac{\partial P^{s+T-r-s-1}_{(1,1)}}{\partial \rho_s} = \frac{T-s-r}{T} \frac{\partial P^{r}_{(1,1)}}{\partial \rho_r} + \cdots + \frac{1}{T} \frac{\partial P^{r+T-r-s-1}_{(1,1)}}{\partial \rho_r}$$

By Lemma A.1, we know that the above is true. Hence (A20) holds and $d \ln |\frac{\partial g}{\partial f'}|$ is exact. We can conclude that (A18) has a solution with the following form,

$$\frac{1}{N} \ln \left| \frac{\partial g}{\partial f'} \right| = \sum_{r=1}^{p} R_r(\rho) + k$$

where k is an arbitrary constant not depending on ρ, $R_1(\rho) = \int h_1(\rho) d\rho_1$ and

$$R_r(\rho) = \int \left(h_r(\rho) - \sum_{j=1}^{r-1} \frac{\partial R_j(\rho)}{\partial \rho_r} \right) d\rho_r \qquad \textit{for} \quad r = 2, \ldots, p.$$

□

A.11. Find the solution for (28) under predetermined regressors

After integrating out f with a flat prior, one can obtain

$$\prod_{i=1}^{N} p(y_i|\theta, z_i, X_i, y_{i,0}) \propto |U|^{\frac{N}{2}} \exp\Bigg[-\frac{1}{2}\sum_{i=1}^{N}\left(Qy_i - Qy_{i\_}\rho - QX_i\beta - \Gamma\tilde{x}_i - \zeta z_i'\alpha\right)$$
$$'U\left(Qy_i - Qy_{i\_}\rho - QX_i\beta - \Gamma\tilde{x}_i - \zeta z_i'\alpha\right)\Bigg] \tag{A21}$$

where $Q = [I_{T-1}, -(\iota_{T-1} - \zeta)]$, $y_{i\_} = (y_{i,0}, y_{i,1}, \ldots, y_{i,T-1})'$ and

$$U = \Omega^{-1} - \frac{\Omega^{-1}(\iota - \zeta)(\iota - \zeta)'\Omega^{-1}}{(\iota - \zeta)'\Omega^{-1}(\iota - \zeta) + (\sigma)^{-2}} = \left(\Omega + \sigma^2(\iota - \zeta)(\iota - \zeta)'\right)^{-1}.$$

Since the last row in $\Sigma_{\epsilon x}$ only contains zeros, one just need to consider its first $T-1$ rows denoted as $\tilde{\Sigma}_{\epsilon x}$. The parameters to be estimated are $\theta = (\alpha, \beta, \rho, \tilde{\Sigma}_{\epsilon x}, \delta, \sigma^2_{f|z}, \sigma^2)$. To estimate them, one can solve (28) or,

$$\begin{aligned}
& d\sum_{i=1}^{N} \ln p(y_i|\theta, z_i, X_i, y_{i,0}) + \sum_{\theta^r} \frac{d\ln\left|\frac{\partial g}{\partial f'}\right|}{d\theta^r} d\theta^r = 0 \\
& s.t. R\, vec(\tilde{\Sigma}_{\epsilon x}) = 0, \sigma^2 > 0, \sigma^2_{f|z} > 0, \\
& \quad \Omega = \sigma^2 I_{T-1} - \Gamma\Sigma_{x|z}\Gamma' + \sigma^2_{f|z}\zeta\zeta' \text{ is positive definite,} \\
& \quad \Sigma_{x|z,f} = \Sigma_{x|z} - \sigma^2_{f|z}\delta\delta' \text{ is positive definite.}
\end{aligned}$$

The matrix R is to impose the restrictions to ensure each row of $\tilde{\Sigma}_{\epsilon x}$ starts with $K \times t$ zeros for $t = 0, 1, \ldots T-2$. Using (42), (43) and (A21), one can rewrite the equation to be solved as

$$\begin{aligned}
0 = & \frac{1}{2}tr(U^{-1}dU) - \frac{1}{2}\sum_{i=1}^{N}[Q(y_i - y_{i_}\rho - X_i\beta) - \Gamma\tilde{x}_i - \zeta z_i'\alpha]'dU[Q(y_i - y_{i_}\rho - X_i\beta) - \Gamma\tilde{x}_i - \zeta z_i'\alpha] + \\
& \sum_{i=1}^{N}[Q(y_{i_}d\rho + X_i d\beta) + d\Gamma\tilde{x}_i + \zeta z_i' d\alpha - d\zeta(y_{iT} - y_{i,T-1}\rho - x_{iT}'\beta - z_i'\alpha)]' \\
& U\left[Q(y_i - y_{i_}\rho - X_i\beta) - \Gamma\tilde{x}_i - \zeta z_i'\alpha\right] - N\left(\frac{(\iota - \zeta)'\Omega^{-1}}{\nu}\right)d\zeta + N h_\rho(\rho, \tilde{\Sigma}_{\epsilon x}, \delta, \sigma^2_{f|z}, \sigma^2)d\rho
\end{aligned} \tag{A22}$$

From the terms only involving $d\alpha$ and $d\beta$, one can obtain

$$\begin{bmatrix}\hat{\beta} \\ \hat{\alpha}\end{bmatrix} = \begin{bmatrix}\sum_{i=1}^{N} X_i'Q'UQX_i & \sum_{i=1}^{N} X_i'Q'U\zeta z_i' \\ \sum_{i=1}^{N} z_i\zeta'UQX_i & \zeta'U\zeta\sum_{i=1}^{N} z_i z_i'\end{bmatrix}^{-1}\begin{bmatrix}\sum_{i=1}^{N} X_i'Q'U(Qy_i - \Gamma\tilde{x}_i) \\ \sum_{i=1}^{N} z_i\zeta'U(Qy_i - \Gamma\tilde{x}_i)\end{bmatrix}.$$

Further note that

$$\begin{aligned}
dU = & U\Big\{\sigma^2 d\zeta(\iota - \zeta)' + \sigma^2(\iota - \zeta)d\zeta' - d\sigma^2[I + (\iota - \zeta)(\iota - \zeta)'] \\
& + d\tilde{\Sigma}_{\epsilon x}\Sigma^{-1}_{x|z,f}\tilde{\Sigma}'_{\epsilon x} + \tilde{\Sigma}_{\epsilon x}d\Sigma^{-1}_{x|z,f}\tilde{\Sigma}'_{\epsilon x} + \tilde{\Sigma}_{\epsilon x}\Sigma^{-1}_{x|z,f}d\tilde{\Sigma}'_{\epsilon x}\Big\}U, \\
d\Sigma^{-1}_{x|z,f} = & \Sigma^{-1}_{x|z,f}\left(d\sigma^2_{f|z}\delta\delta' + (\sigma_{f|z})^2 d\delta\delta' + \sigma^2_{f|z}\delta d\delta'\right)\Sigma^{-1}_{x|z,f}, \\
d\zeta = & d\tilde{\Sigma}_{\epsilon x}\Sigma^{-1}_{x|z,f}\delta + \tilde{\Sigma}_{\epsilon x}d\Sigma^{-1}_{x|z,f}\delta + \tilde{\Sigma}_{\epsilon x}\Sigma^{-1}_{x|z,f}d\delta, \\
d\Gamma = & d\tilde{\Sigma}_{\epsilon x}\Sigma^{-1}_{x|z,f} + \tilde{\Sigma}_{\epsilon x}d\Sigma^{-1}_{x|z,f}.
\end{aligned}$$

Substituting the above into (A22) and setting the terms before each differential 0, we can estimate $\tilde{\Sigma}_{\epsilon x}, \delta, \rho, \sigma^2_{f|z}$ and σ^2 by solving the following equations:

$$\frac{M_{R'}}{N} vec\Bigg\{ U \sum_{i=1}^{N} Qe_i \left[\tilde{x}_i - (y_{i,T} - y_{i,T-1}\rho - x'_{i,T}\beta - \alpha' z_i)\delta' \right]' \Sigma^{-1}_{x|z,f} - (\nu\Omega)^{-1}(\iota - \zeta)\delta' \Sigma^{-1}_{x|z,f}$$

$$+ U\left[U^{-1} - \frac{1}{N}\sum_{i=1}^{N} Qe_i e_i' Q' \right] U \left[\tilde{\Sigma}_{\epsilon x} + \sigma^2 (\iota - \zeta)\delta' \right] \Sigma^{-1}_{x|z,f} \Bigg\} = 0,$$

$$\Sigma^{-1}_{x|z,f} \tilde{\Sigma}'_{\epsilon x} U \left[U^{-1} - \frac{1}{N}\sum_{i=1}^{N} Qe_i e_i' Q \right] U \left[\sigma^2_{f|z} \tilde{\Sigma}_{\epsilon x} \Sigma^{-1}_{x|z,f}\delta + \sigma^2 (1 + \sigma^2_{f|z}\delta' \Sigma^{-1}_{x|z,f}\delta)(\iota - \zeta) \right]$$

$$+ \sigma^2 (\sigma_{f|z})^2 \Sigma^{-1}_{x|z,f}\delta(\iota - \zeta)' U \left[U^{-1} - \frac{1}{N}\sum_{i=1}^{N} Qe_i e_i' Q' \right] U \tilde{\Sigma}_{\epsilon x} \Sigma^{-1}_{x|z,f}\delta$$

$$+ \frac{\sigma^2_{f|z}}{N} \Sigma^{-1}_{x|z,f} \left[\sum_{i=1}^{N} \tilde{x}_i e_i' Q' U \tilde{\Sigma}_{\epsilon x} + (\sum_{i=1}^{N} \tilde{x}_i e_i' Q' U \tilde{\Sigma}_{\epsilon x})' \right] \Sigma^{-1}_{x|z,f}\delta$$

$$- \frac{\sigma^2_{f|z}}{N} \Sigma^{-1}_{x|z,f}\delta \sum_{i=1}^{N} (y_{i,T} - y_{i,T-1}\rho - x'_{i,T}\beta - \alpha' z_i) e_i' Q' U \tilde{\Sigma}_{\epsilon x} \Sigma^{-1}_{x|z,f}\delta$$

$$- \frac{\sigma^2_{f|z}\delta' \Sigma^{-1}_{x|z,f}\delta + 1}{N} \Sigma^{-1}_{x|z,f} \tilde{\Sigma}'_{\epsilon x} U \sum_{i=1}^{N} (y_{i,T} - y_{i,T-1}\rho - x'_{i,T}\beta - \alpha' z_i) Qe_i - \Sigma^{-1}_{x|z,f} \tilde{\Sigma}'_{\epsilon x} (\nu\Omega)^{-1}(\iota - \zeta)$$

$$- \frac{\sigma^2_{f|z}}{\nu} \left[\delta' \Sigma^{-1}_{x|z,f}\delta \Sigma^{-1}_{x|z,f} \tilde{\Sigma}'_{\epsilon x} \Omega^{-1}(\iota - \zeta) + (\iota - \zeta)' \Omega^{-1} \tilde{\Sigma}_{\epsilon x} \Sigma^{-1}_{x|z,f}\delta \Sigma^{-1}_{x|z,f}\delta \right] = 0,$$

$$\frac{1}{N}\sum_{i=1}^{N} y'_{i\_} Q' U Qe_i + h_\rho(\rho, \tilde{\Sigma}_{\epsilon x}, \delta, \sigma^2_{f|z}, \sigma^2) = 0$$

$$\left[\sigma^2 \delta' \Sigma^{-1}_{x|z,f}\delta(\iota - \zeta) + \frac{1}{2} \tilde{\Sigma}_{\epsilon x} \Sigma^{-1}_{x|z,f}\delta \right]' U \left[U^{-1} - \frac{1}{N}\sum_{i=1}^{N} Qe_i e_i' Q' \right] U \tilde{\Sigma}_{\epsilon x} \Sigma^{-1}_{x|z,f}\delta$$

$$- \delta' \Sigma^{-1}_{x|z,f}\delta\delta' \Sigma^{-1}_{x|z,f} \tilde{\Sigma}'_{\epsilon x} \left[(\nu\Omega)^{-1}(\iota - \zeta) + \frac{U}{N}\sum_{i=1}^{N} \left(y_{i,T} - y_{i,T-1}\rho - x'_{i,T}\beta - \alpha' z_i \right) Qe_i \right]$$

$$+ \frac{1}{N}\delta' \Sigma^{-1}_{x|z,f} \tilde{\Sigma}'_{\epsilon x} U \sum_{i=1}^{N} Qe_i \tilde{x}_i' \Sigma^{-1}_{x|z,f}\delta = 0,$$

$$- \frac{1}{2} tr \left\{ U \left(U^{-1} - \frac{1}{N}\sum_{i=1}^{N} Qe_i e_i' Q' \right) U \left[I + (\iota - \zeta)(\iota - \zeta)' \right] \right\} = 0.$$

where $Qe_i = Qy_i - Qy_{i\_}\rho - Qy_{i\_}\rho - QX_i\beta - \Gamma\tilde{x}_i - \zeta\alpha' z_i$ and $M_{R'} = I - R'(RR')^{-1}R$.

A.12. Details of generating the explanatory variable

In all cases, there is only one explanatory, x_{it}, which is always stationary and can be strictly exogenous or predetermined. The generating processes of y_{it} are given in (44) for stationary case and (45) for non-stationary case. Below are the processes followed by x_{it}.

- When y_{it} is stationary and x_{it} is predetermined, the process for x_{it} is

$$x_{it} = 0.3x_{i,t-1} + 0.3y_{i,t-1} + 0.3y_{i,t-2} + u_{it}.$$

- When y_{it} is non-stationary and x_{it} is predetermined, x_{it} is from

$$x_{it} = 0.3f_i + 0.51x_{i,t-1} + 0.3\epsilon_{i,t-1} + u_{it}.$$

- When y_{it} is stationary and x_{it} is exogenous, the process for x_{it} is

$$x_{it} = 0.3f_i + 0.39x_{i,t-1} + u_{it}.$$

- When y_{it} is non-stationary and x_{it} is exogenous, the process for x_{it} is

$$x_{it} = 0.3f_i + 0.51x_{i,t-1} + u_{it}.$$

  where $u_{it} \sim i.i.d.N(0, 1)$ and is independent from ϵ_{it}.

Acknowledgments

We would like to extend our heartfelt thanks to the editors, associate editors, and anonymous referees for their thoughtful and constructive feedback, which has greatly enriched and strengthened this paper.

References

Arellano, M., Bonhomme, S. (2009). Robust priors in nonlinear panel data models. *Econometrica* 77:489–536.
Arellano, M., Hahn, J. (2007). Understanding bias in nonlinear panel models: some recent developments. In: Blundell, R., Newey, W., Persson, T., eds., *Advances in Economics and Econometrics, Ninth World Congress*. Cambridge: Cambridge University Press.
Barndorff-Nielsen, O. E., Cox, D. R. (1994). *Inference and Asymptotics*. Boca Raton, FL: Chapman & Hall /CRC.
Bartolucci, F., Bellio, R., Salvan, A., Sartori, N. (2016). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews* 35(7):1271–1289. doi: 10.1080/07474938.2014.975642
Bester, C. A., Hansen, C. (2009). A penalty function approach to bias reduction in nonlinear panel models with fixed effects. *Journal of Business & Economic Statistics* 27(2):131–148. doi: 10.1198/jbes.2009.0012
Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys* 2:107–144.
Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions Volume I*. Norman, OK: Kendrick Press, Inc.
Carro, J. M. (2007). Estimating dynamic panel data discrete choice models with fixed effects. *Journal of Econometrics* 140(2):503–528. doi: 10.1016/j.jeconom.2006.07.023
Cox, D. R., Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 49(1):1–18. doi: 10.1111/j.2517-6161.1987.tb01422.x
Dhaene, G., Jochmans, K. (2016). Likelihood inference in an Autoregression with fixed effects. *Econometric Theory* 32(5):1178–1215. doi: 10.1017/S0266466615000146
Fernandez-Val, I. (2009). Fixed effects estimation of structural parameters and marginal effects in panel probit models. *Journal of Econometrics* 150:71–85.
Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80(1):27–38. doi: 10.1093/biomet/80.1.27
Gourieroux, C., Monfort, A., Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* 52(3):681. doi: 10.2307/1913471
Greene, W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal* 7(1):98–119. doi: 10.1111/j.1368-423X.2004.00123.x
Hahn, J., Kuersteiner, G. (2011). Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory* 27(6):1152–1191. doi: 10.1017/S0266466611000028
Hahn, J., Newey, W. (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72(4):1295–1319. doi: 10.1111/j.1468-0262.2004.00533.x

Heckman, J. J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic processs. In *Structural Analysis of Discrete Data with Econometric Applictions*, MIT Press, pp. 179–195.

Kass, R. E., L., Tierney, J. B., Kadane., (1990). The validity of posterior expansions based on laplace's method. In: Geisser, S., Hodges, J. S., Press, S. J., Zellner, A., eds. *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George a Barnard*. Amsterdam, Netherlands: North-Holland.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics* 95(2):391–413. doi: 10.1016/S0304-4076(99)00044-5

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies* 69(3):647–666. doi: 10.1111/1467-937X.t01-1-00025

Li, G. (2015). Consistency in estimation and model selection of dynamic panel data models with fixed effects. *Econometrics* 3(3):494–524. doi: 10.3390/econometrics3030494

Moral-Benito, E. (2013). Likelihood-based estimation of dynamic panels with predetermined regressors. *Journal of Business & Economic Statistics* 31(4):451–472. doi: 10.1080/07350015.2013.818003

Moreira, M. J. (2009). A maximum likelihood method for the incidental parameter problem. *The Annals of Statistics* 37:3660–3696.

Neyman, J., Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16(1):1 doi: 10.2307/1914288

Ord, K., Arnold, S., Stuart, A. (1999). *Kendall's Advanced Theory of Statistics: Classical Inference and the Linear Model 2*. London: Arnold.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* 90(3):533–549. doi: 10.1093/biomet/90.3.533

Severini, T. A. (2000). *Inference and Asymptotics*. Oxford: Oxford University Press.

Sweeting, T. J. (1995). A bayesian approach to approximate conditional inference. *Biometrika* 82(1):25–36. doi: 10.1093/biomet/82.1.25

Tanaka, K. (1983). Asymptotic expansions associated with the AR(1) model with unknown mean. *Econometrica* 51(4):1221 doi: 10.2307/1912060

White, H. (2001). *Asymptotic Theory for Econometricians*. Cambridge, MA: Academic Press

Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics* 182(1):226–234. doi: 10.1016/j.jeconom.2014.04.020

Woutersen, T. (2003). Robustness against incidental parameters. Unpublished Manuscript, The University of Arizona, Tucson, AZ.