Split-panel jackknife estimation of fixed-effect models

Geert Dhaene, Koen Jochmans

Split-panel jackknife estimation of fixed-effect models

Geert Dhaene* Koen Jochmans†
K.U. Leuven CORE

This version: January 2010

Abstract

We propose a jackknife for reducing the order of the bias of maximum likelihood estimates of nonlinear dynamic fixed-effect panel models. In its simplest form, the half-panel jackknife, the estimator is just $2\widehat{\theta} - \overline{\theta}_{1/2}$, where $\widehat{\theta}$ is the MLE from the full panel and $\overline{\theta}_{1/2}$ is the average of the two half-panel MLEs, each using $T/2$ time periods and all N cross-sectional units. This estimator eliminates the first-order bias of $\widehat{\theta}$. The order of the bias is further reduced if two partitions of the panel are used, for example, two half-panels and three 1/3-panels, and the corresponding MLEs. On further partitioning the panel, any order of bias reduction can be achieved. The split-panel jackknife estimators are asymptotically normal, centered at the true value, with variance equal to that of the MLE under asymptotics where T is allowed to grow slowly with N. In analogous fashion, the split-panel jackknife reduces the bias of the profile likelihood and the bias of marginal-effect estimates. Simulations in fixed-effect dynamic discrete-choice models with small T show that the split-panel jackknife effectively reduces the bias and mean-squared error of the MLE, and yields confidence intervals with much better coverage.

JEL classification: C13, C14, C22, C23
Keywords: jackknife, asymptotic bias correction, dynamic panel data, fixed effects.

*Address: K.U. Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium. Tel. +32 16 326798; Fax +32 16 326796; E-mail: `geert.dhaene@econ.kuleuven.be`.

†Address: U.C. Louvain, Center for Operations Research and Econometrics, Voie du Roman Pays 34, B-1348 Louvain-la-Neuve, Belgium. Tel. +32 10 474329; E-mail: `koen.jochmans@uclouvain.be`.

1 Introduction

Fixed effects in panel data models in general cause the maximum likelihood estimator of the parameters of interest to be inconsistent if the length of the panel, T, remains fixed while the number of cross-sectional units, N, grows large. This is the incidental parameter problem, first noted by Neyman and Scott (1948). Lancaster (2000) gives a review. For certain models, it is possible to separate the estimation of the fixed effects from inference about the common parameters, for example, by conditioning on a sufficient statistic, as in logit models (Rasch, 1961; Andersen, 1970; Chamberlain, 1980), or by using moment conditions that are free of fixed effects, as in the dynamic linear model (Anderson and Hsiao, 1981, 1982).[1] However, these approaches are model specific and give no direct guidance to estimating average effects. A general solution to the incidental parameter problem does not exist and seems impossible due to the lack of point identification in certain models (Chamberlain, 2010; Honoré and Tamer, 2006) or singularity of the information matrix (Hahn, 2001; Magnac, 2004).

A recent strand in the literature, aiming at greater generality, looks for estimators that reduce the inconsistency (or asymptotic bias) of the MLE by an order of magnitude, that is, from $O(T^{-1})$ down to $O(T^{-2})$.[2] Lancaster (2000, 2002), Woutersen (2002), Arellano (2003), and Arellano and Bonhomme (2009) argued that a suitably modified likelihood or score function approximately separates the estimation of the fixed effects from the estimation of the common parameters. Arellano and Hahn (2006, 2007) and Bester and Hansen (2009) proposed modifications to the profile likelihood that remove its leading bias term relative to a target likelihood that is free of fixed effects. Hahn and Newey (2004) and Hahn and Kuersteiner (2004) derived the leading term in an expansion of the bias of the MLE as T grows and constructed an estimator of it. Alternatively, as shown by Hahn and Newey (2004) for i.i.d. panel data, an estimate of the leading bias term may also be obtained by applying a delete-one panel jackknife, extending Quenouille (1956). All these approaches lead to estimates that are first-order unbiased and, unlike the MLE, have an asymptotic distribution that is correctly centered as N and T grow at the same rate.[3]

We propose a split-panel jackknife (SPJ) for reducing the bias of the MLE in dynamic

[1] See Chamberlain (1984) and Arellano and Honoré (2001) for surveys.

[2] Arellano and Hahn (2007) give an overview of currently existing results.

[3] There has been a similar development in the statistics literature on inference in the presence of nuisance parameters. See, for example, Cox and Reid (1987) and Sweeting (1987) on the role of information orthogonality, and Firth (1993), Severini (2000), Li, Lindsay, and Waterman (2003), Sartori (2003), and Pace and Salvan (2006) on modified profile likelihoods and score functions.

models, adapting ideas of Quenouille (1949), who was interested in reducing the bias of estimates from time series, to the panel setting. The jackknife exploits the property that the bias of the MLE can be expanded in powers of T^{-1}.[4] By comparing the MLE from the full sample with ML estimates computed from subsamples, an estimate of the bias up to a chosen order is obtained. In a panel setting with fixed effects, the subsamples are subpanels with fewer observations along the time dimension. In its simplest form, the panel is split into two non-overlapping half-panels, each with $T/2$ time periods and all N cross-sectional units. If $\overline{\theta}_{1/2}$ is the average of the ML estimates corresponding to the half-panels and $\widehat{\theta}$ is the MLE from the full panel, then the bias of $\widehat{\theta}$ is roughly half of the bias of $\overline{\theta}_{1/2}$ and, therefore, is estimated by $\overline{\theta}_{1/2} - \widehat{\theta}$. Subtracting this estimate from $\widehat{\theta}$ gives the half-panel jackknife estimator, $2\widehat{\theta} - \overline{\theta}_{1/2}$, which is first-order unbiased. Its asymptotic distribution is normal, correctly centered, and has variance equal to that of the MLE, if $N/T^3 \to 0$ as $N, T \to \infty$. By partitioning the panel further, an appropriately weighted average of subpanel ML estimates admits any order of bias reduction without inflating the asymptotic variance. An h-order SPJ estimator has bias $O(T^{-h-1})$ and is asymptotically normal and efficient if $N/T^{2h+1} \to 0$ as $N, T \to \infty$. We give an asymptotic characterization of the transformation that the SPJ induces on the remaining bias terms, similar to the characterization of Adams, Gray, and Watkins (1971) in a cross-sectional framework with i.i.d. data, and derive a simple rule for selecting the partitions that minimize the impact of jackknifing on the remaining bias. For standard errors and confidence sets, we propose to use the bootstrap or the jackknife where resampling or subsampling occurs over the cross-sectional units.[5] The SPJ may be applied in analogous fashion to bias-correct the likelihood. The maximizer of the jackknifed profile loglikelihood inherits the bias reduction induced on the likelihood and, under asymptotics where $N, T \to \infty$ and T is allowed to grow slowly with N, is equivalent to the SPJ applied to the MLE. Similarly, the SPJ yields bias-corrected estimates of average marginal and other effects, where the averaging is over the fixed effects.

In Section 2 we introduce the panel model of interest and some notation. The SPJ correction to the MLE is developed in Section 3. Sections 4 and 5 deal with corrections to the profile likelihood and to average effect estimates, respectively. The results of a Monte Carlo application to dynamic discrete-choice models are reported in Section 6.

---

[4] Miller (1974) contains a review on the jackknife. See also Shao and Tu (1995).

[5] In a cross-sectional framework, Brillinger (1964) and Reeds (1978) showed that the estimate obtained by jackknifing the MLE has the same asymptotic distribution as the MLE and that the jackknife estimate of variance is consistent.

Section 7 concludes. An appendix contains the proofs.

2 Framework and assumptions

Let the data be $z_{it} \equiv (y_{it}, x_{it})$, where $i = 1, ..., N$ and $t = 1, ..., T$. We make the following assumption.

Assumption 1. *For all i, the processes $z_{it} = (y_{it}, x_{it})$ are stationary and alpha mixing, with mixing coefficients that decrease at an exponential rate, and are independent across i. The conditional density of y_{it}, given x_{it} (relative to some dominating measure), is $f(y_{it}|x_{it}; \theta_0, \alpha_{i0})$, where (θ_0, α_{i0}) is the unique maximizer of $\mathbb{E} \log f(y_{it}|x_{it}; \theta, \alpha_i)$ over the Euclidean parameter space $\Theta \times \mathcal{A}$ and is interior to it.*

Assumption 1 allows x_{it} to contain lagged values of y_{it} and of covariates, thus accommodating dynamic panel data. It also allows feedback of y on covariates. The variables y_{it}, x_{it} and the parameters θ, α_i may be vectors. Our interest lies in estimating θ_0.

Let $f_{it}(\theta, \alpha_i) \equiv f(y_{it}|x_{it}; \theta, \alpha_i)$. The MLE of θ_0 is

$$\widehat{\theta} \equiv \arg\max_{\theta} \widehat{l}(\theta), \qquad \widehat{l}(\theta) \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \log f_{it}(\theta, \widehat{\alpha}_i(\theta)),$$

where $\widehat{\alpha}_i(\theta) \equiv \arg\max_{\alpha_i} \frac{1}{T} \sum_{t=1}^{T} \log f_{it}(\theta, \alpha_i)$ and $\widehat{l}(\theta)$ is the profile loglikelihood, normalized by the number of observations. For fixed T, $\widehat{\theta}$ is generally inconsistent for θ_0, that is, $\theta_T \equiv \text{plim}_{N \to \infty} \widehat{\theta} \neq \theta_0$ (Neyman and Scott, 1948) due to the presence of incidental parameters, $\alpha_1, ..., \alpha_N$. This is because, under regularity conditions,

$$\theta_T = \arg\max_{\theta} l_T(\theta), \qquad l_T(\theta) \equiv \overline{\mathbb{E}} \log f_{it}(\theta, \widehat{\alpha}_i(\theta)),$$

where $\overline{\mathbb{E}}(\cdot)$ denotes $\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(\cdot)$, while

$$\theta_0 = \arg\max_{\theta} l_0(\theta), \qquad l_0(\theta) \equiv \overline{\mathbb{E}} \log f_{it}(\theta, \alpha_i(\theta)),$$

where $\alpha_i(\theta) \equiv \arg\max_{\alpha_i} \mathbb{E} \log f_{it}(\theta, \alpha_i)$. Therefore, with $\widehat{\alpha}_i(\theta) \neq \alpha_i(\theta)$, the maximands $l_T(\theta)$ and $l_0(\theta)$ are different and so, in general, are their maximizers.

We make the following assumptions about the asymptotic bias, $\theta_T - \theta_0$, and about the large N, T distribution of $\widehat{\theta}$. Let $s_{it}(\theta) \equiv \partial \log f_{it}(\theta, \alpha_i(\theta))/\partial\theta$, $s_{it} \equiv s_{it}(\theta_0)$, and $\Omega \equiv [\overline{\mathbb{E}}(s_{it} s_{it}')]^{-1}$.

Assumption 2. θ_T *exists and, as* $T \to \infty$,

$$\theta_T = \theta_0 + \frac{B_1}{T} + \frac{B_2}{T^2} + ... + \frac{B_k}{T^k} + o\left(T^{-k}\right), \tag{2.1}$$

where k *is a positive integer and* $B_1, ..., B_k$ *are constants.*

Assumption 3. Ω *exists and, as* $N, T \to \infty$,

$$\widehat{\theta} - \theta_T = \frac{\Omega}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} + o_p\big((NT)^{-1/2}\big). \tag{2.2}$$

Assumption 2 is the key requirement for the split-panel jackknife to reduce the asymptotic bias of $\widehat{\theta}$ from $O(T^{-1})$ to a smaller order. Hahn and Kuersteiner (2004) give conditions under which (2.1) holds for $k = 1$. Assumptions 1–3 imply that

$$\sqrt{NT}(\widehat{\theta} - \theta_T) \xrightarrow{d} \mathcal{N}(0, \Omega) \quad \text{as } N, T \to \infty.$$

For example, if $N/T \to \kappa$, then $\sqrt{NT}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(B_1\sqrt{\kappa}, \Omega)$. Thus, while $\widehat{\theta}$ is consistent for θ_0 as $N, T \to \infty$, it is asymptotically incorrectly centered when T grows at the same rate as N or more slowly (Hahn and Kuersteiner, 2004).[6] Under Assumptions 1–3, jackknifing $\widehat{\theta}$ will asymptotically re-center the estimate at θ_0, even when T grows slowly with N.

One may view the asymptotic bias of $\widehat{\theta}$ as resulting from the inconsistency of $\widehat{l}(\theta)$ for $l_0(\theta)$, i.e. $l_T(\theta) = \text{plim}_{N\to\infty}\widehat{l}(\theta) \neq l_0(\theta)$, which suggests that one may also jackknife $\widehat{l}(\theta)$ instead of $\widehat{\theta}$. We make the following assumptions, analogous to Assumptions 2 and 3, about the asymptotic bias $l_T(\theta) - l_0(\theta)$ and about the large N, T distribution of the profile score, $\widehat{s}(\theta) \equiv \partial\widehat{l}(\theta)/\partial\theta$. Let $s_T(\theta) \equiv \text{plim}_{N\to\infty}\widehat{s}(\theta)$, $s_0(\theta) \equiv \partial l_0(\theta)/\partial\theta$, and $\Omega(\theta) \equiv [\overline{\mathbb{E}} \sum_{j=-\infty}^{\infty} \text{Cov}(s_{it}(\theta), s_{it-j}(\theta))]^{-1}$. Note that $s_0(\theta_0) = 0$ and that $\Omega(\theta_0) = \Omega = -[\partial s_0(\theta)/\partial\theta']_{\theta=\theta_0}$.

Assumption 4. *There is a neighborhood of* θ_0 *where* $l_T(\theta)$ *exists and, as* $T \to \infty$,

$$l_T(\theta) = l_0(\theta) + \frac{C_1(\theta)}{T} + \frac{C_2(\theta)}{T^2} + ... + \frac{C_k(\theta)}{T^k} + o\left(T^{-k}\right), \tag{2.3}$$

where k *is a positive integer and* $C_1, ..., C_k$ *are functions, each with a bounded derivative.*

[6] This also occurs in dynamic linear models (Hahn and Kuersteiner, 2002; Alvarez and Arellano, 2003) and in nonlinear models with i.i.d. data (Hahn and Newey, 2004).

Assumption 5. *There is a neighborhood of θ_0 where $\Omega(\theta)$ and $s_T(\theta)$ exist and, as $N, T \to \infty$,*

$$\widehat{s}(\theta) - s_T(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(s_{it}(\theta) - s_0(\theta)\big) + o_p\big((NT)^{-1/2}\big).$$

Arellano and Hahn (2006) give conditions under which (2.3) holds for $k = 1$. Under Assumptions 1 and 4–5, jackknifing $\widehat{l}(\theta)$ will asymptotically re-center its maximizer at θ_0, even when T grows slowly with N.

3 Bias correction of the MLE

We derive the split-panel jackknife estimator as a weighted average of the MLE and MLEs defined by subpanels. A subpanel is defined as a proper subset $S \subsetneq \{1, ..., T\}$ such that the elements of S are consecutive integers and $|S| \geq T_{\min}$, where $T_{\min}$ is the least T for which θ_T exists.[7] The MLE corresponding to a subpanel S is

$$\widehat{\theta}_S \equiv \arg\max_{\theta} \widehat{l}_S(\theta) \ , \qquad \widehat{l}_S(\theta) \equiv \frac{1}{N|S|} \sum_{i=1}^{N} \sum_{t \in S} \log f_{it}(\theta, \widehat{\alpha}_{iS}(\theta)),$$

where $\widehat{\alpha}_{iS}(\theta) \equiv \arg\max_{\alpha_i} \frac{1}{|S|} \sum_{t \in S} \log f_{it}(\theta, \alpha_i)$.

Since subpanels by their definition preserve the time-series structure of the full panel, stationarity implies $\text{plim}_{N\to\infty} \widehat{\theta}_S = \theta_{|S|}$ and, as $|S| \to \infty$, $\theta_{|S|}$can be expanded as in (2.1), with $|S|$ replacing T. By taking a suitable weighted average of $\widehat{\theta}$ and MLEs defined by subpanels, one or more of the leading terms of the bias of $\widehat{\theta}$ can be eliminated. There are many different ways to achieve this, and, as a result, a whole range of bias-corrected estimators is obtained.

The SPJ can be seen as transforming $B_1, ..., B_k$ into $0, ..., 0, B'_{h+1}, ..., B'_k$, thus (i) eliminating the first h terms of the bias of $\widehat{\theta}$ and (ii) transforming the higher-order bias terms that are not eliminated. We derive this transformation explicitly. Naturally, the SPJ estimators can be classified by the order of bias correction achieved, h. Estimators with the same h can be further classified as to whether or not the large N, T variance is inflated (and, if so, by how much) and by the implied coefficients of the higher-order bias terms that are not eliminated, $B'_{h+1}, ..., B'_k$. These coefficients are always larger (in absolute value) than $B_{h+1}, ..., B_k$, respectively. For SPJ estimators that do not inflate

[7] We use $|A|$ to denote the cardinality of A when A is a set, the absolute value when A is a real number, and the determinant when A is a square matrix.

the large N,T variance, there is a lower bound on $B'_{h+1},...,B'_k$. This bound increases rapidly with h and is attained under a very simple rule for selecting the subpanels. When one is prepared to accept variance inflation, there exist SPJ estimators that reduce the bias further either by further increasing the order of bias correction, h, or by reducing $B'_{h+1},...,B'_k$. Although the variance inflation may be substantial, so may be the additional bias reduction, especially when T is very small and hence the bias of $\widehat{\theta}$ is likely to be large.

Our estimators are motivated by asymptotic arguments that involve both $N \to \infty$ and $T \to \infty$. We have no theoretical results for fixed T. Nevertheless, because our asymptotics will allow T to grow very slowly with N, they are intended to give a reasonable approximation to the properties of the estimators in applications where T may be (though need not be) small compared to N.

3.1 First-order bias correction

Suppose for a moment that T is even. Partition $\{1,...,T\}$ into two half-panels, $S_1 \equiv \{1,...,T/2\}$ and $S_2 \equiv \{T/2+1,...,T\}$, and let $\overline{\theta}_{1/2} \equiv \frac{1}{2}(\widehat{\theta}_{S_1}+\widehat{\theta}_{S_2})$. Clearly, $\text{plim}_{N\to\infty}\overline{\theta}_{1/2} = \theta_{T/2}$ and so, the half-panel jackknife estimator

$$\widehat{\theta}_{1/2} \equiv 2\widehat{\theta} - \overline{\theta}_{1/2} \tag{3.1}$$

has an asymptotic bias

$$\begin{aligned}\text{plim}_{N\to\infty}\widehat{\theta}_{1/2} - \theta_0 &= -2\frac{B_2}{T^2} - 6\frac{B_3}{T^3} - ... - (2^k-2)\frac{B_k}{T^k} + o\left(T^{-k}\right) \\ &= O(T^{-2})\end{aligned}$$

if (2.1) holds with $k \geq 2$. That is, $\widehat{\theta}_{1/2}$ is a first-order bias-corrected estimator of θ_0; it is free of bias up to $O(T^{-2})$. Assumptions 1 and 3 imply

$$\sqrt{NT}\begin{pmatrix} \widehat{\theta} - \theta_T \\ \overline{\theta}_{1/2} - \theta_{T/2} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega & \Omega \\ \Omega & \Omega \end{pmatrix}\right) \qquad \text{as } N,T\to\infty,$$

and, in turn, $\sqrt{NT}(\widehat{\theta}_{1/2} - 2\theta_T + \theta_{T/2}) \xrightarrow{d} \mathcal{N}(0,\Omega)$. Thus, $\widehat{\theta}_{1/2}$ has the same large N,T variance as $\widehat{\theta}$. Under asymptotics where $N,T \to \infty$ and $N/T^3 \to 0$, we have $\sqrt{NT}(2\theta_T - \theta_{T/2} - \theta_0) = \sqrt{NT}O(T^{-2}) \to 0$. Therefore,

$$\sqrt{NT}(\widehat{\theta}_{1/2} - \theta_0) \xrightarrow{d} \mathcal{N}(0,\Omega) \qquad \text{as } N,T\to\infty \text{ and } N/T^3 \to 0.$$

Thus, $\widehat{\theta}_{1/2}$ is asymptotically correctly centered at θ_0 whenever T grows faster than $N^{1/3}$.[8] These properties carry over to a more general class of SPJ estimators.

Let $g \geq 2$ be an integer. For $T \geq gT_{\min}$, let $\mathcal{S} \equiv \{S_1, ..., S_g\}$ be a collection of non-overlapping subpanels such that $\cup_{S\in\mathcal{S}} S = \{1, ..., T\}$ and the sequence $\min_{S\in\mathcal{S}} |S|/T$ is bounded away from zero. Define the SPJ estimator

$$\widehat{\theta}_{\mathcal{S}} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}_{\mathcal{S}}, \qquad \overline{\theta}_{\mathcal{S}} \equiv \sum_{S\in\mathcal{S}} \frac{|S|}{T}\widehat{\theta}_S.$$

Theorem 1. *Let Assumptions 1 and 2 hold. If $k = 1$, then* $\text{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}} = \theta_0 + o\left(T^{-1}\right)$. *If $k \geq 2$, then*

$$\text{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}} = \theta_0 + \frac{B_2'}{T^2} + \frac{B_3'}{T^3} + ... + \frac{B_k'}{T^k} + o\left(T^{-k}\right)$$

where

$$B_j' \equiv \frac{g - T^{j-1}\sum_{S\in\mathcal{S}} |S|^{1-j}}{g-1} B_j = O(1),$$

$\text{sign}(B_j') = -\text{sign}(B_j)$, *and* $|B_j'| \geq |B_j| \sum_{m=1}^{j-1} g^m$. *If Assumptions 1, 2, and 3 hold for some $k \geq 2$, then*

$$\sqrt{NT}(\widehat{\theta}_{\mathcal{S}} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \textit{as } N, T \to \infty \textit{ and } N/T^3 \to 0. \tag{3.2}$$

Theorem 1 requires the collection of subpanels, $\mathcal{S}$, to be a partition of $\{1, ..., T\}$. This condition is not needed for bias correction but is required for not inflating the large N, T variance of $\widehat{\theta}_{\mathcal{S}}$. When (in an asymptotically non-negligible sense) $\mathcal{S}$ does not cover $\{1, ..., T\}$ or when some subpanels intersect, the large N, T variance of $\widehat{\theta}_{\mathcal{S}}$ (with $\overline{\theta}_{\mathcal{S}}$ suitably redefined as $\sum_{S\in\ \mathcal{S}} |S|\widehat{\theta}_S / \sum_{S\in\mathcal{S}} |S|$) exceeds Ω. We will state this more precisely in Subsection 3.3.

While $\widehat{\theta}_{\mathcal{S}}$ eliminates the first-order bias of $\widehat{\theta}$ without increasing the large N, T variance, this happens at the cost of increasing the magnitude of the higher-order bias terms, since $\sum_{m=1}^{j-1} g^m > 1$ for $j \geq 2$. For a given g, any higher-order bias coefficient, B_j', is minimized (in absolute value) if and only if $\sum_{S\in\mathcal{S}} |S|^{1-j}$ is minimized. This occurs if and only if the subpanels $S \in \mathcal{S}$ have approximately equal length, that is, for all

[8] Clearly, under the same asymptotics, $2\widehat{\theta} - \widehat{\theta}_S$ is also correctly centered and free of bias up to $O(T^{-2})$ for any subpanel S with $|S| = T/2$. This estimator bears resemblance to an estimator suggested by Hu (2002, pp. 2512–2513), although she seems to be using an estimate based on $T/2$ randomly selected time periods instead of $\widehat{\theta}_S$. Whether such an approach reduces the asymptotic bias of $\widehat{\theta}$ is unclear, since the random selection alters the dependency structure of the data and hence may affect the coefficients in (2.1). Further, unlike $\widehat{\theta}_{1/2}$, an estimator of the form $2\widehat{\theta} - \widehat{\theta}_S$ has inflated large N, T variance, equal to 2Ω.

$S \in \mathcal{S}$, either $|S| = \lfloor T/g \rfloor$ or $|S| = \lceil T/g \rceil$. Thus, within the class $\widehat{\theta}_{\mathcal{S}}$ with given g, the equal-length SPJ estimator

$$\widehat{\theta}_{1/g} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}_{1/g}, \qquad \overline{\theta}_{1/g} \equiv \sum_{S \in \mathcal{S}} \frac{|S|}{T}\widehat{\theta}_S,$$
$$\text{where } |S| = \lfloor T/g \rfloor \text{ or } |S| = \lceil T/g \rceil \text{ for all } S \in \mathcal{S},$$

minimizes all higher-order bias terms. The subscript $1/g$ indicates that each subpanel is approximately a fraction $1/g$ of the full panel.[9] It follows from Theorem 1 that $\widehat{\theta}_{1/g}$ has second-order bias $-gB_2/T^2$. Hence, within the class $\widehat{\theta}_{\mathcal{S}}$, all higher-order bias terms are minimized by the half-panel jackknife estimator,

$$\widehat{\theta}_{1/2} \equiv 2\widehat{\theta} - \overline{\theta}_{1/2}, \qquad \overline{\theta}_{1/2} \equiv \frac{|S_1|}{T}\widehat{\theta}_{S_1} + \frac{|S_2|}{T}\widehat{\theta}_{S_2},$$
$$\text{where } S_1 = \{1, ..., \lfloor T/2 \rfloor\} \text{ or } S_1 = \{1, ..., \lceil T/2 \rceil\}, \text{ and } S_2 = \{1, ..., T\} \backslash S_1,$$

which slightly generalizes (3.1) in that T is allowed to be odd. This provides a theoretical justification for using the half-panel jackknife—of course, within the confines of the class $\widehat{\theta}_{\mathcal{S}}$. As will be shown in Subsection 3.2, the higher-order bias terms of $\widehat{\theta}_{1/2}$ can be further eliminated up to some order determined by T.

The half-panel jackknife estimator is very easy to compute. All that is needed are three maximum likelihood estimates. When N is large, as is often the case in microeconometric panels, a computationally efficient algorithm for obtaining maximum likelihood estimates will exploit the sparsity of the Hessian matrix, as, for example, in Hall (1978). Furthermore, once $\widehat{\theta}$ and $\widehat{\alpha}_1, ..., \widehat{\alpha}_N$ are computed, they are good starting values for computing $\widehat{\theta}_{S_1}$ and $\widehat{\alpha}_{1S_1}, ..., \widehat{\alpha}_{NS_1}$; in turn, $2\widehat{\theta} - \widehat{\theta}_{S_1}$ and $2\widehat{\alpha}_1 - \widehat{\alpha}_{1S_1}, ..., 2\widehat{\alpha}_N - \widehat{\alpha}_{NS_1}$ are good starting values for computing $\widehat{\theta}_{S_2}$ and $\widehat{\alpha}_{1S_2}, ..., \widehat{\alpha}_{NS_2}$.[10]

The half-panel jackknife may be seen as an automatic way of estimating and removing the first-order bias of $\widehat{\theta}$. Unlike the analytically bias-corrected estimator of Hahn and Kuersteiner (2004), it avoids the need of a plug-in estimate for estimating the leading term of $\theta_T - \theta_0$, and with it the need to choose smoothing parameters. Both estimators have zero first-order bias and have the same limiting distribution as $N, T \to \infty$ and $N/T^3 \to 0$. However, their second-order biases will generally be different. While the SPJ inflates the magnitude of all remaining bias terms, the analytical bias correction

[9] When T is not divisible by g, there are several ways to split the panel into g approximately equal-length subpanels, all yielding estimators $\widehat{\theta}_{1/g}$ with the same bias. Averaging $\widehat{\theta}_{1/g}$ over all possible choices of $\mathcal{S}$ removes any arbitrariness arising from a particular choice of $\mathcal{S}$ but does not affect the bias.

[10] For sufficiently large T, the Newton-Raphson algorithm, starting from the values mentioned, converges in one iteration.

alters those terms through the use of the MLE as a plug-in estimate. Presumably, the use of an iterative procedure, as in Hahn and Newey (2004), will leave the second-order bias term unaffected.

The jackknife, as a method for bias reduction, originated in the seminal work of Quenouille (1949, 1956). Quenouille (1949) argued that, in a time series context, the first-order bias of the sample autocorrelation coefficient, say $\widehat{\rho}$, is eliminated by using two half-series to form $2\widehat{\rho} - \overline{\rho}_{1/2}$, in obvious notation. Quenouille (1956) observed that, when an estimator T_n, based on n i.i.d. observations, has bias $O(n^{-1})$, the estimator $nT_n - (n-1)\overline{T}_{n-1}$ (later termed the delete-one jackknife estimator), where $\overline{T}_{n-1}$ is the average of the n statistics T_{n-1}, often has bias $O(n^{-2})$. The half-panel jackknife is the natural extension of Quenouille's (1949) half-series jackknife to fixed-effect panel data, just as Hahn and Newey's (2004) panel jackknife extends Quenouille's (1956) delete-one jackknife to fixed-effect panel data that are i.i.d. across time.

The jackknife is a much more powerful bias-reducing device in fixed-effect panels than in a single time series or single cross-section framework, where it was originally used. If $N/T \to \infty$, the squared bias dominates in the asymptotic mean squared error of $\widehat{\theta}$, which is $O(N^{-1}T^{-1}) + O(T^{-2})$. The jackknife, operating on the dominant term, reduces the asymptotic MSE to $O(N^{-1}T^{-1}) + O(T^{-4})$. By contrast, in a time series or a cross-section setting, it leaves the asymptotic MSE unchanged at $O(T^{-1})$ or $O(N^{-1})$.

3.2 Higher-order bias correction

As shown, a suitable linear combination of the MLE and a weighted average of non-overlapping subpanel MLEs removes the first-order bias of the MLE without large N, T variance inflation. The use of two half-panels gives the least second- and higher-order bias terms. Continuing the arguments, we find that they yield second- and higher-order bias corrections.[11] A suitable linear combination of the MLE and two weighted averages of MLEs, each one associated with a collection of non-overlapping subpanels, removes the first- and second-order bias without large N, T variance inflation. The use of two half-panels and three 1/3-panels gives the least third- and higher-order bias terms. And so on.

To see how the SPJ can eliminate the second-order bias of $\widehat{\theta}$, suppose for a moment that T is divisible by 2 and 3, and let $G = \{2, 3\}$. Then the estimator $\widehat{\theta}_{1/G} = (1 + a_{1/2} +$

[11] With i.i.d. cross-sectional data, Quenouille (1956) already noted that a second-order bias correction is obtained by re-applying the delete-one jackknife, with slight modification, to $nT_n - (n-1)\overline{T}_{n-1}$. The idea was later generalized to higher-order corrections by Schucany, Gray, and Owen (1971).

$a_{1/3})\widehat{\theta} - a_{1/2}\overline{\theta}_{1/2} - a_{1/3}\overline{\theta}_{1/3}$ has zero first- and second-order biases if $a_{1/2}$ and $a_{1/3}$ satisfy

$$\left(\frac{1 + a_{1/2} + a_{1/3}}{T} - \frac{a_{1/2}}{T/2} - \frac{a_{1/3}}{T/3}\right) B_1 = 0,$$
$$\left(\frac{1 + a_{1/2} + a_{1/3}}{T^2} - \frac{a_{1/2}}{(T/2)^2} - \frac{a_{1/3}}{(T/3)^2}\right) B_2 = 0,$$

regardless of B_1 and B_2. This gives $a_{1/2} = 3$, $a_{1/3} = -1$, and

$$\widehat{\theta}_{1/G} \equiv 3\widehat{\theta} - 3\overline{\theta}_{1/2} + \overline{\theta}_{1/3}, \qquad G = \{2, 3\}.$$

$\widehat{\theta}_{1/G}$ has an asymptotic bias

$$\begin{aligned} \text{plim}_{N\to\infty}\widehat{\theta}_{1/G} - \theta_0 &= 6\frac{B_3}{T^3} + 36\frac{B_4}{T^4} + ... + (3 - 3 \times 2^k + 3^k)\frac{B_k}{T^k} + o\left(T^{-k}\right) \\ &= O(T^{-3}) \end{aligned}$$

if (2.1) holds with $k \geq 3$. Further, by the arguments given earlier,

$$\sqrt{NT}(\widehat{\theta}_{1/G} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \text{as } N, T \to \infty \text{ and } N/T^5 \to 0.$$

That is, $\widehat{\theta}_{1/G}$ has the same large N, T variance as Ω and is asymptotically correctly centered at θ_0 when T grows faster than $N^{1/5}$.

We now introduce SPJ estimators that remove the bias terms of $\widehat{\theta}$ up to order $h \leq k$, without inflating the large N, T variance. Let $G \equiv \{g_1, ..., g_h\}$ be a non-empty set of integers, with $2 \leq g_1 < ... < g_h$. For $T \geq g_h T_{\min}$ and each $g \in G$, let $\mathcal{S}_g$ be a collection of g non-overlapping subpanels such that $\cup_{S\in\mathcal{S}_g} S = \{1, ..., T\}$ and, for all $S \in \mathcal{S}_g$, $|S| = \lfloor T/g \rfloor$ or $|S| = \lceil T/g \rceil$. Let A be the $h \times h$ matrix with elements

$$A_{rs} \equiv \sum_{S\in\mathcal{S}_{g_s}} \left(\frac{T}{|S|}\right)^{r-1}, \qquad r, s = 1, ..., h,$$

and let a_{1/g_r} be the rth element of $(1 - \iota' A^{-1}\iota)^{-1} A^{-1}\iota$, where ι is the $h \times 1$ summation vector. Define the SPJ estimator

$$\widehat{\theta}_{1/G} \equiv \left(1 + \sum_{g\in G} a_{1/g}\right)\widehat{\theta} - \sum_{g\in G} a_{1/g}\overline{\theta}_{1/g}, \qquad \overline{\theta}_{1/g} \equiv \sum_{S\in\mathcal{S}_g} \frac{|S|}{T}\widehat{\theta}_S. \tag{3.3}$$

To describe the higher-order bias of $\widehat{\theta}_{1/G}$, let

$$b_j(G) \equiv (-1)^h g_1...g_h \sum_{\substack{k_1,...,k_h\geq 0 \\ k_1+...+k_h\leq j-h-1}} g_1^{k_1}...g_h^{k_h}, \qquad j = 1, 2, ..., \tag{3.4}$$

with the standard convention that empty sums and products are 0 and 1, respectively, so that $b_j(G) = 0$ for $j \leq h = |G|$, and $b_j(\varnothing) = 1$ for all $j \geq 1$.

Theorem 2. *(i) Let Assumptions 1 and 2 hold for some $k \geq h$. If $k = h$, then* $\text{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + o\left(T^{-h}\right)$. *If $k > h$, then*

$$\text{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + \frac{B'_{h+1}(G)}{T^{h+1}} + \ldots + \frac{B'_k(G)}{T^k} + o\left(T^{-k}\right) \tag{3.5}$$

where $B'_j(G) = b_j(G)B_j + O(T^{-1})$. (ii) If Assumptions 1, 2, and 3 hold for some $k > h$, then

$$\sqrt{NT}(\widehat{\theta}_{1/G} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \text{as } N, T \to \infty \text{ and } N/T^{2h+1} \to 0. \tag{3.6}$$

As the result shows, the SPJ estimator defined in (3.3) eliminates the low-order bias terms of the MLE without large N, T variance inflation and, hence, is correctly centered at θ_0 under slow T asymptotics. However, this occurs at the cost of increasing the higher-order bias terms that are not eliminated, roughly by a factor of $b_j(G)$.[12] For given h, the factors $b_j(G)$ all have the same sign, regardless of G and j. The sign alternates in h. For any given h, $|b_j(G)|$ is minimal for all $j > h$ if and only if $G = \{2, 3, ..., h+1\}$. This choice of G is the SPJ that we tend to recommend because (i) it eliminates the low-order bias terms of the MLE at the least possible increase of the higher-order bias terms and (ii) it attains the Cramér-Rao bound under slow T asymptotics, whereas the MLE only attains this bound when T grows faster than N, i.e., when $N/T \to 0$. Even with this optimal choice of G, the factors $b_j(G)$ increase rapidly as h grows. Table 1 gives the first few values. The elements on the main diagonal of the table are the leading non-zero bias factors, $b_{h+1}(G) = (-1)^h(h+1)!$, $h = 0, 1, \ldots$

Table 1: Higher-order bias factors of the SPJ

	$b_1(\cdot)$	$b_2(\cdot)$	$b_3(\cdot)$	$b_4(\cdot)$	$b_5(\cdot)$
$\widehat{\theta}$	1	1	1	1	1
$\widehat{\theta}_{1/2}$	0	−2	−6	−14	−30
$\widehat{\theta}_{1/\{2,3\}}$	0	0	6	36	150
$\widehat{\theta}_{1/\{2,...,4\}}$	0	0	0	−24	−240
$\widehat{\theta}_{1/\{2,...,5\}}$	0	0	0	0	120

Regarding the choice of h, extending the arguments given above would suggest choosing $h = \lfloor T/T_{\min}\rfloor - 1$, which is the largest value for which the SPJ estimator (3.3) is defined. However, we do not recommend this choice except, perhaps, when T is relatively small, for at least three reasons. First, in the asymptotics, we kept h fixed while

[12]When T increases in multiples of the least common multiple of $g_1, ..., g_l$, (3.5) holds with $B'_j(G)$ exactly equal to $b_j(G)B_j$.

$T \to \infty$, so we have no justification for letting h grow large with T. Second, as $T \to \infty$, the bias of $\widehat{\theta}$ (and that of any fixed-h SPJ estimator) vanishes, and so does the gain in terms of (higher-order) bias reduction. Third, the choice of h should also be guided by variance considerations. Our analysis yields the same first-order asymptotic variance for all SPJ estimators 3.3) and the MLE. However, just as the SPJ affects the bias terms of all orders, it also affects the higher-order variance terms. To shed light on this question, higher-order asymptotic variance calculations would be required, which are beyond the scope of this paper.

3.3 Bias correction with overlapping subpanels

The SPJ estimator defined in (3.3) uses h collections of non-overlapping subpanels to eliminate the first h bias terms of $\widehat{\theta}$. The same can be achieved by using collections of overlapping subpanels. Subpanel overlap has two main effects: (i) it permits the higher-order bias coefficients $B'_{h+1}, ..., B'_k$ to be substantially smaller than is otherwise possible; (ii) it increases the large N, T variance. Thus, a trade-off between high-order bias reduction and large N, T variance minimization arises (though see the remark at the end of this subsection).

To fix ideas, suppose T is divisible by g, a rational number strictly between 1 and 2. Let S_1 and S_2 be subpanels such that $S_1 \cup S_2 = \{1, ..., T\}$ and $|S_1| = |S_2| = T/g$. Consider the SPJ estimator

$$\widehat{\theta}_{1/g} \equiv \frac{g}{g-1}\widehat{\theta} - \frac{1}{g-1}\overline{\theta}_{1/g}, \qquad \overline{\theta}_{1/g} \equiv \frac{1}{2}(\widehat{\theta}_{S_1} + \widehat{\theta}_{S_2}), \tag{3.7}$$

where, as before, the subscript $1/g$ indicates that each subpanel uses a fraction $1/g$ of the full panel. This estimator has asymptotic bias

$$\text{plim}_{N\to\infty}\widehat{\theta}_{1/g} - \theta_0 = -g\frac{B_2}{T^2} - g(1+g)\frac{B_3}{T^3} - ... - g(1+g+...+g^{k-2})\frac{B_k}{T^k} + o\left(T^{-k}\right).$$

Each term of this bias is smaller (in magnitude) than the corresponding bias term of $\widehat{\theta}_{1/2}$. As g decreases from 2 to 1, the overlap between the subpanels increases and the higher-order bias coefficients $B'_j = -g\sum_{h=0}^{j-2} g^h B_j$ decrease to $(1-j)B_j$ (in magnitude). Regarding the large N, T variance, a simple calculation gives

$$\sqrt{NT}\begin{pmatrix} \widehat{\theta} - \theta_T \\ \overline{\theta}_{1/g} - \theta_{T/g} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega & \Omega \\ \Omega & \frac{g(3-g)}{2}\Omega \end{pmatrix}\right) \qquad \text{as } N, T \to \infty,$$

and hence

$$\sqrt{NT}\left(\widehat{\theta}_{1/g} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g}{2(g-1)}\Omega\right) \qquad \text{as } N, T \to \infty \text{ and } N/T^3 \to 0.$$

As g decreases from 2 to 1, the large N, T variance of $\widehat{\theta}_{1/g}$ increases from Ω to ∞.

We now consider SPJ estimators where there may be collections of non-overlapping subpanels and collections of two overlapping subpanels. Let $0 \leq o \leq h$, $1 \leq h$, and $G \equiv \{g_1, ..., g_h\}$, where $1 < g_1 < ... < g_o < 2 \leq g_{o+1} < ... < g_h$ and $g_{o+1}, ..., g_h$ are integers. For $T \geq g_h T_{\min}$ and T large enough such that $\lceil T/g \rceil \neq \lceil T/g' \rceil$ for all distinct $g, g' \in G$, let, for each $g \in G$, $\mathcal{S}_g$ be a collection of subpanels such that (i) $\cup_{S \in \mathcal{S}_g} S = \{1, ..., T\}$; (ii) if $g < 2$, then $\mathcal{S}_g$ consists of two subpanels, each with $\lceil T/g \rceil$ elements; (iii) if $g \geq 2$, $\mathcal{S}_g$ consists of g non-overlapping subpanels and, for all $S \in \mathcal{S}_g$, $|S| = \lfloor T/g \rfloor$ or $|S| = \lceil T/g \rceil$. Define the SPJ estimator

$$\widehat{\theta}_{1/G} \equiv \left(1 + \sum_{g \in G} a_{1/g}\right) \widehat{\theta} - \sum_{g \in G} a_{1/g} \bar{\theta}_{1/g}, \qquad \bar{\theta}_{1/g} \equiv \sum_{S \in \mathcal{S}_g} \frac{|S|}{\sum_{S \in \mathcal{S}_g} |S|} \widehat{\theta}_S, \tag{3.8}$$

where a_{1/g_r} is the rth element of $(1 - \iota' A^{-1} \iota)^{-1} A^{-1} \iota$ and A is the $h \times h$ matrix with elements

$$A_{rs} \equiv \frac{\sum_{S \in \mathcal{S}_{g_s}} (T/|S|)^{r-1}}{\sum_{S \in \mathcal{S}_{g_s}} |S|/T}, \qquad r, s = 1, ..., h. \tag{3.9}$$

Note that, when $o = 0$, $\widehat{\theta}_{1/G}$ reduces to the SPJ estimator given in (3.3). Let $b(G)$ be as in (3.4), and let

$$d_T(G) \equiv 1 + (1 - \iota' A^{-1} \iota)^{-2} \iota' A'^{-1} \Gamma A^{-1} \iota,$$

where Γ is the symmetric $h \times h$ matrix whose (r, s)th element, for $r \leq s$, is

$$\Gamma_{rs} \equiv \begin{cases} \frac{1}{2} (A_{1r} - 1)(2 - A_{1s}) & \text{if } s \leq o, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 3. *With $\widehat{\theta}_{1/G}$ redefined by (3.8) and (3.9), part (i) of Theorem 2 continues to hold and, if Assumptions 1, 2, and 3 hold for some $k > h$, then*

$$\sqrt{\frac{NT}{d_T(G)}} (\widehat{\theta}_{1/G} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \text{as } N, T \to \infty \text{ and } N/T^{2h+1} \to 0, \tag{3.10}$$

and $d(G) \equiv \lim_{T \to \infty} d_T(G) \geq 1$, with equality if and only if $o = 0$.

Overlapping subpanels allow $|b_j(G)|$ to be much smaller than is possible with collections of non-overlapping subpanels because $|b_j(G)|$ increases rapidly in all $g \in G$. For the same reason, the optimal choice of $g_{o+1}, ..., g_h$, from the perspective of minimizing the higher-order bias terms, is $2, ..., h-o+1$. However, with overlapping subpanels, the large N, T variance inflation factor, $d_T(G)$, increases very rapidly with both the number of collections of overlapping subpanels, o, and the number of collections of non-overlapping subpanels, $h - o$. To illustrate the variance inflation, Table 2 gives the minimum value of $d(G)$ when there are up to two collections of overlapping subpanels ($o = 1, 2$) and up to three collections of non-overlapping subpanels ($h - o = 1, 2, 3$), the latter with $g_{o+1}, ..., g_h$ set equal to 2 to up to 4. The minimum of $d(G)$ is computed over $g_1, ..., g_o$, given o, and the minimizers, $g_1^*, ..., g_o^*$, are also given in the table. The minimum of $d(G)$ increases very rapidly in o, so in practice one would hardly ever consider using more than one collection of overlapping subpanels in combination with collections of non-overlapping subpanels. The case without non-overlapping subpanels ($o = h$) is not treated in the table (where it would correspond to $G_2 = \varnothing$) because, for given $o = h$, the least value of $d(G)$ is reached as g_o approaches 2, implying that, for $o = h = 1, 2, 3$, we have $\inf_G d(G) = 1, 9, 124.5$, respectively.

Table 2: Variance inflation factors of the SPJ with overlapping subpanels

G_1^*	G_2		
	$\{2\}$	$\{2,3\}$	$\{2,3,4\}$
	9	30.0	66.1
$\{g_1^*\}$	$\{1.5\}$	$\{1.36\}$	$\{1.30\}$
	124.5	440.2	1039.7
$\{g_1^*, g_2^*\}$	$\{1.20, 1.84\}$	$\{1.15, 1.77\}$	$\{1.13, 1.72\}$

Note: The entries are the minimal variance inflation factor, $d(G_1^* \cup G_2)$, and the corresponding $G_1^* = \arg\min_{G_1 : \max G_1 < 2} d(G_1 \cup G_2)$, given G_2 and $o = |G_1|$.

Consideration of the variance inflation factor, while based on large N, T arguments that may be inaccurate when T is small, suggests that the SPJ with overlapping subpanels should only be used in applications where N is very large and there is a great need for bias reduction, for example, when T is very small. Note, however, that, when $T_{\min} < T < 2T_{\min}$, the SPJ can only be applied if the subpanels overlap.

Subpanel overlap causes large N, T variance inflation because the time periods, t, receive unequal weights in those $\overline{\theta}_{1/g}$ where $1 < g < 2$. In principle, it is possible to prevent variance inflation by adding to $\overline{\theta}_{1/g}$ a term, with zero probability limit, that equalizes those weights. As an example, take $g = 3/2$ and suppose T is a multiple of 3 and $T \geq 3T_{\min}$. Then

$$\overline{\theta}_{2/3} = \frac{1}{2}(\widehat{\theta}_{1:2} + \widehat{\theta}_{2:3}),$$

where $\widehat{\theta}_{1:2}$ and $\widehat{\theta}_{2:3}$ use the first two-thirds and the last two-thirds of the time periods, respectively. Now consider

$$\widetilde{\theta}_{2/3} \equiv \frac{1}{2}(\widehat{\theta}_{1:2} + \widehat{\theta}_{2:3}) + \frac{1}{12}(\widehat{\theta}_{1:1} - 2\widehat{\theta}_{2:2} + \widehat{\theta}_{3:3}),$$

where each t receives a weight $1/T$ and $\text{plim}_{N\to\infty}\widetilde{\theta}_{2/3} = \text{plim}_{N\to\infty}\overline{\theta}_{2/3}$ because the second term of $\widetilde{\theta}_{2/3}$ has zero probability limit. Hence, replacing $\overline{\theta}_{2/3}$ with $\widetilde{\theta}_{2/3}$ in $\widehat{\theta}_{1/G}$, with unchanged weights $a_{1/g}$, $g \in G$, will leave the asymptotic bias unaffected but will reduce the large N, T variance. It is possible, for any $T \geq 2T_{\min}$ and any $g \in (1, 2)$ that divides T, to find $\widetilde{\theta}_{1/g}$, similar to $\widetilde{\theta}_{2/3}$, such that each t receives a weight $1/T$ and $\text{plim}_{N\to\infty}\widetilde{\theta}_{1/g} = \text{plim}_{N\to\infty}\overline{\theta}_{1/g}$. However, the weights associated with certain subpanel MLEs in the zero plim term may become large, especially when g is close to 1, similar to the weights of the delete-one estimates in the ordinary jackknife. In simulations with small T, we found that this may substantially increase the variance, so we leave the idea for future work.

3.4 Variance estimation and confidence sets

Let $\widehat{\theta}_{1/G}$ be an SPJ estimator of the form (3.8) and suppose Assumptions 1 to 3 hold for some $k > h$. Consider asymptotics where $N, T \to \infty$ and $N/T^{2h+1} \to 0$, so that $\widehat{\theta}_{1/G}$ is asymptotically normal and centered at θ_0. For estimating $\text{Var}(\widehat{\theta}_{1/G})$ (assuming it exists) and for constructing confidence sets for θ_0, we propose to use the bootstrap, where the i's are resampled, or the delete-one-i jackknife.[13] We assume that $\alpha_1, ..., \alpha_N$ are i.i.d. random draws from some distribution, thus implying that $z_1, ..., z_N$, where $z_i \equiv (z_{i1}, ..., z_{iT})'$, are i.i.d. random vectors. The bootstrap and jackknife are then essentially the same as in the case of i.i.d. cross-sectional data.

Write the original panel as $z \equiv (z_1, ..., z_N)$ and the SPJ estimator as $\widehat{\theta}_{1/G}(z)$. Define a bootstrap panel as a draw $\widetilde{z} \equiv (z_{d_1}, ..., z_{d_N})$ where $d_1, ..., d_N$ are i.i.d. random draws from $\{1, ..., N\}$. Thus, the columns of $\widetilde{z}$ are columns of z drawn with replacement. The bootstrap distribution of $\widehat{\theta}_{1/G}$ is the distribution of $\widehat{\theta}_{1/G}(\widetilde{z})$, given z. Its variance is a consistent estimate of $\text{Var}(\widehat{\theta}_{1/G})$ (in the sense that the ratio of estimate to estimand converges weakly to 1) and α-probability minimum-volume ellipsoids are confidence sets with asymptotic coverage α.

[13] Kapetanios (2008) suggested this version of the bootstrap for fixed-effect linear panel data models. It is perhaps also worth noting that, under our asymptotics, the usual estimate of the asymptotic variance of the MLE is also a consistent estimate of the asymptotic variance of the SPJ estimator (at least when there is no subpanel overlap; otherwise, the estimate is to be multiplied by $d_T(G)$).

Let z_{-i} be obtained from z on deleting its i-th column. The N quantities $N\widehat{\theta}_{1/G}(z) - (N-1)\widehat{\theta}_{1/G}(z_{-i})$ can then be viewed as pseudo-values, in the sense of Tukey (1958), associated with $\widehat{\theta}_{1/G}$. The pseudo-values are nearly independent across i and have nearly the same distribution as $\widehat{\theta}_{1/G}$. The jackknife distribution of $\widehat{\theta}_{1/G}$ is the uniform distribution on the set of pseudo-values, given z. It can be used in the same way as the bootstrap distribution to deliver a consistent estimate of $\mathrm{Var}(\widehat{\theta}_{1/G})$ and asymptotically correct α-confidence sets.

4 Bias correction of the likelihood

In Section 3 the SPJ was used to remove the low-order bias terms of $\widehat{\theta}$. It can also be used, in a completely analogous fashion, to remove the low-order bias terms of the profile loglikelihood, $\widehat{l}(\theta)$.

Let $T'_{\min}$ be the least T for which $l_T(\theta)$ exists and is non-constant.[14] To remove the first-order bias term of $l_T(\theta)$ using half-panels, let $T \geq 2T'_{\min}$, suppose T is even, let $S_1 \equiv \{1, ..., T/2\}$ and $S_2 \equiv \{T/2+1, ..., T\}$, and define the half-panel jackknife profile loglikelihood as

$$\widehat{l}_{1/2}(\theta) \equiv 2\widehat{l}(\theta) - \bar{l}_{1/2}(\theta), \qquad \bar{l}_{1/2}(\theta) \equiv \frac{1}{2}\left(\widehat{l}_{S_1}(\theta) + \widehat{l}_{S_2}(\theta)\right).$$

Then

$$\begin{aligned}\mathrm{plim}_{N\to\infty}\widehat{l}_{1/2}(\theta) - l_0(\theta) &= -2\frac{C_2(\theta)}{T^2} - 6\frac{C_3(\theta)}{T^3} - ... - (2^k-2)\frac{C_k(\theta)}{T^k} + o\left(T^{-k}\right) \\ &= O(T^{-2})\end{aligned}$$

if (2.3) holds with $k \geq 2$. Thus, $\widehat{l}_{1/2}(\theta)$ is free of bias up to $O(T^{-2})$, and so is the corresponding SPJ estimator,

$$\dot{\theta}_{1/2} = \arg\max_{\theta} \widehat{l}_{1/2}(\theta).$$

Let $\bar{s}_{1/2}(\theta) \equiv \partial\bar{l}_{1/2}(\theta)/\partial\theta$ and $\widehat{s}_{1/2}(\theta) \equiv \partial\widehat{l}_{1/2}(\theta)/\partial\theta$. Assumptions 1 and 5 imply that, in a neighborhood around θ_0,

$$\sqrt{NT}\begin{pmatrix} \widehat{s}(\theta) - s_T(\theta) \\ \bar{s}_{1/2}(\theta) - s_{T/2}(\theta) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega(\theta)^{-1} & \Omega(\theta)^{-1} \\ \Omega(\theta)^{-1} & \Omega(\theta)^{-1} \end{pmatrix}\right) \quad \text{as } N, T \to \infty,$$

[14] The values $T_{\min}$ and $T'_{\min}$ may differ. This occurs, for example, in dynamic binary models. We return to this point in Section 6.

so $\sqrt{NT}(\widehat{s}_{1/2}(\theta) - s_0(\theta)) \xrightarrow{d} \mathcal{N}(0, \Omega(\theta)^{-1})$ as $N, T \to \infty$ and $N/T^3 \to 0$. Hence, because $\widehat{s}_{1/2}(\dot{\theta}_{1/2}) = 0$ with probability approaching 1, $\sqrt{NT}(\dot{\theta}_{1/2} - \theta_0)$ is asymptotically normal and centered at 0. Expanding $\widehat{s}_{1/2}(\dot{\theta}_{1/2}) = 0$ gives

$$0 = \sqrt{NT}\widehat{s}_{1/2}(\theta_0) + \sqrt{NT}\widehat{h}_{1/2}(\theta_0)(\dot{\theta}_{1/2} - \theta_0) + o_p(1)$$

where $\widehat{h}_{1/2}(\theta) \equiv \partial\widehat{s}_{1/2}(\theta)/\partial\theta' = \partial s_0(\theta)/\partial\theta' + o_p(1)$. So $\widehat{h}_{1/2}(\theta_0) = -\Omega^{-1} + o_p(1)$ and

$$\sqrt{NT}(\dot{\theta}_{1/2} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \text{as } N, T \to \infty \text{ and } N/T^3 \to 0.$$

Under asymptotics where $N, T \to \infty$ and $N/T^3 \to 0$, $\widehat{\theta}_{1/2}$ and $\dot{\theta}_{1/2}$ are efficient, so they must be asymptotically equivalent, i.e., $\sqrt{NT}(\widehat{\theta}_{1/2} - \dot{\theta}_{1/2}) \xrightarrow{p} 0$.

The half-panel SPJ provides an automatic way of correcting the bias of the profile likelihood, $\widehat{l}(\theta)$. Analytically bias-corrected profile likelihoods were proposed by Arellano and Hahn (2006) and Bester and Hansen (2009). Relative to $\widehat{l}(\theta)$, all methods give an improved approximation to the target likelihood, $l_0(\theta)$, by removing the first-order term of $l_T(\theta) - l_0(\theta)$. More generally, define the SPJ profile loglikelihood by analogy to (3.8) as

$$\widehat{l}_{1/G}(\theta) \equiv \left(1 + \sum_{g \in G} a_{1/g}\right)\widehat{l}(\theta) - \sum_{g \in G} a_{1/g}\bar{l}_{1/g}(\theta), \qquad \bar{l}_{1/g}(\theta) \equiv \sum_{S \in \mathcal{S}_g} \frac{|S|}{\sum_{S \in \mathcal{S}_g}|S|}\widehat{l}_S(\theta),$$

where $G = \{g_1, ..., g_h\}$ and the collections of subpanels $\mathcal{S}_g$, the matrix A, and the scalars $a_{1/g}$ are as in Subsection 3.3. Then, if Assumptions 1 and 4 hold for some $k \geq h$, in a neighborhood of θ_0, we have $\text{plim}_{N\to\infty}\widehat{l}_{1/G}(\theta) = l_0(\theta) + o(T^{-h})$ and, if $k > h$,

$$\text{plim}_{N\to\infty}\widehat{l}_{1/G}(\theta) = l_0(\theta) + \frac{C'_{h+1}(\theta, G)}{T^{h+1}} + ... + \frac{C'_k(\theta, G)}{T^k} + o\left(T^{-k}\right)$$

where $C'_j(\theta, G) = b_j(G)C_j(\theta) + O(T^{-1})$. The SPJ estimator associated with $\widehat{l}_{1/G}$ is

$$\dot{\theta}_{1/G} = \arg\max_{\theta} \widehat{l}_{1/G}(\theta),$$

and is free of bias up to $o(T^{-h})$ if $k \geq h$, and up to $O(T^{-h-1})$ if $k > h$. If Assumptions 1, 4, and 5 hold for some $k > h$, then

$$\sqrt{\frac{NT}{d_T(G)}}(\dot{\theta}_{1/G} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \text{as } N, T \to \infty \text{ and } N/T^{2h+1} \to 0,$$

and $\dot{\theta}_{1/G}$ and $\widehat{\theta}_{1/G}$ are asymptotically equivalent as $N, T \to \infty$ and $N/T^{2h+1} \to 0$.

Note that $\dot{\theta}_{1/G}$ is equivariant under one-to-one transformations of θ, while $\widehat{\theta}_{1/G}$ is not. Note, also, that applying the SPJ to the profile likelihood, $\widehat{l}(\theta)$, is identical to applying the SPJ to the profile score, $\widehat{s}(\theta)$; that is, the resulting SPJ estimators of θ_0 are the same.

Jackknifing the profile loglikelihood readily extends to unbalanced data. Suppose an unbalanced panel is formed as the union of J independent balanced panels of dimensions $N_j \times T_j$, $j = 1, ..., J$. Let $w_j = N_j T_j / \sum_j N_j T_j$ and let all ratios $N_j/N_{j'}$ and $T_j/T_{j'}$ be fixed as $\sum_j N_j$ or $\sum_j T_j$ grows. The MLE is $\widehat{\theta} = \arg\max_\theta \sum_j w_j \widehat{l}_j(\theta)$, where $\widehat{l}_j(\theta)$ is the normalized profile loglikelihood from the jth balanced panel. Jackknifing $\widehat{l}_j(\theta)$ for each j as described above eliminates the low-order bias terms of $\widehat{l}_j(\theta)$ and hence of $\sum_j w_j \widehat{l}_j(\theta)$. The estimator maximizing the jackknifed version of $\sum_j w_j \widehat{l}_j(\theta)$ is asymptotically normal and correctly centered as $\sum_j N_j, \sum_j T_j \to \infty$ with $\sum_j T_j$ allowed to increase slowly with $\sum_j N_j$. By contrast, it is not immediately clear how one should jackknife $\widehat{\theta}$ directly because, in general, $\text{plim}_{\sum_j N_j \to \infty} \widehat{\theta} \neq \sum_j w_j \theta_{T_j}$. One may, however, jackknife the asymptotically equivalent estimator $\sum_j w_j \widehat{\theta}_j$, where $\widehat{\theta}_j$ is the MLE from the jth balanced panel, by jackknifing $\widehat{\theta}_j$ for each j.

5 Bias correction for average effects

Suppose we are interested in the quantity μ_0 defined by the moment condition

$$\overline{\mathbb{E}} q(\mu_0, w, z_{it}, \theta_0, \alpha_{i0}) = 0$$

for some known function $q(\cdot)$ and chosen value w, where $\dim q = \dim \mu_0$. This includes averages and quantiles of marginal or non-marginal effects at fixed or observed covariate values. For example, in the probit model $\Pr[y_{it} = 1 | x_{it}] = \Phi(\alpha_{i0} + \theta_0 x_{it})$, one may be interested in the average effect (on the choice probabilities) of changing x_{it} from w_1 to w_2, $\mu_0 \equiv \overline{\mathbb{E}}(\Phi(\alpha_{i0} + \theta_0 w_2) - \Phi(\alpha_{i0} + \theta_0 w_1))$, or in the average marginal effect of x_{it} at observed values, $\mu_0 \equiv \theta_0 \overline{\mathbb{E}} \phi(\alpha_{i0} + \theta_0 x_{it})$.[15]

The SPJ readily extends to this setting. A natural estimator for μ_0 is the value $\widehat{\mu}$ that solves

$$\widehat{q}(\widehat{\mu}, \widehat{\theta}) = 0, \qquad \widehat{q}(\mu, \theta) \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} q(\mu, w, z_{it}, \theta, \widehat{\alpha}_i(\theta)).$$

[15] For some models the bias of average effect estimates may be negligibly small because the biases may nearly cancel out by averaging over the cross-sectional units. See Hahn and Newey (2004) for examples and Fernández-Val (2009) for theoretical results in the static probit model. This small bias property does not generally hold, however, for models with dynamics (Fernández-Val, 2009).

Whenever $\widehat{\mu}$ has an asymptotic bias that can be expanded in integer powers of T^{-1}, it can be corrected for bias by jackknifing $(\widehat{\mu}, \widehat{\theta})$ or $(\widehat{q}, \widehat{l})$.

6 Simulations for dynamic discrete-choice models

We chose fixed-effect dynamic probit and logit models as a test case for the SPJ. When T is small, the MLE in these models is heavily biased.[16] Here, we present results for the following probit models:

$$\begin{aligned} &\text{AR(1):} && y_{it} = 1(\alpha_{i0} + \rho_0 y_{it-1} + \varepsilon_{it} \geq 0), && \varepsilon_{it} \sim \mathcal{N}(0,1), \\ &\text{ARX(1):} && y_{it} = 1(\alpha_{i0} + \rho_0 y_{it-1} + \beta_0 x_{it} + \varepsilon_{it} \geq 0), && \varepsilon_{it} \sim \mathcal{N}(0,1). \end{aligned}$$

The data were generated with $\alpha_i \sim \mathcal{N}(0,1)$ and $x_{it} = .5x_{it-1} + u_{it}$, where $u_{it} \sim \mathcal{N}(0,1)$. For both models, the initial observations, y_{i0} and x_{i0}, were drawn from their stationary distributions. We set $N = 500$; $T = 6, 9, 12, 18$; $\rho_0 = .5, 1$; $\beta_0 = .5$; and ran 10,000 Monte Carlo replications at each design point, with all random variables redrawn in each replication.

We estimated the common parameter, $\theta_0 = \rho_0$ in the AR(1) and $\theta_0 = (\rho_0, \beta_0)'$ in the ARX(1), by the MLE, $\widehat{\theta}$; the analytically bias-corrected estimators of Hahn and Kuersteiner (2004) and Arellano and Hahn (2006), $\widehat{\theta}_{\text{HK}}$ and $\dot{\theta}_{\text{AH}}$;[17] and the SPJ estimators $\widehat{\theta}_{1/2}$, $\widehat{\theta}_{1/\{2,3\}}$, $\dot{\theta}_{1/2}$, and $\dot{\theta}_{1/\{2,3\}}$, with one exception.[18] When $T = 6$, $\widehat{\theta}_{1/\{2,3\}}$ is not defined because $T_{\min} = 3$; it was replaced by $\widehat{\theta}_{1/\{3/2,2\}}$. This is the only case where subpanels overlap (in a given collection) and the corresponding figures in the tables below are in italics.[19] When $T = 9$, the subpanels $\{1, ..., 5\}$ and $\{6, ..., 9\}$ were used for the SPJ estimators. The other values of T are divisible by 3/2, 2, and 3, so they always allow equal-length subpanels in each collection. There is a positive probability that $\widehat{\theta}$ is indeterminate or infinite, which implies non-existence of moments and possible numerical difficulties in computing the MLE. In the supplementary material to this paper we characterize the data for which the MLE is indeterminate or infinite. In the simulations, when $\widehat{\theta}$ was either indeterminate or infinite, the data set was discarded and

[16]See, for example, the Monte Carlo results reported by Greene (2004) and Carro (2007).

[17]The bandwidth was set equal to 1 for $\widehat{\theta}_{\text{HK}}$ and $\dot{\theta}_{\text{AH}}$, and $\dot{\theta}_{\text{AH}}$ was implemented with the determinant-based approach and Bartlett weights.

[18]Although it was not developed for dependent data, we also experimented with the delete-one panel jackknife. Simulations indicate that its bias is $O(T^{-1})$, like the bias of the MLE.

[19]Interestingly, $T'_{\min} = 2$, so the likelihood can be jackknifed using subpanels of length 2 (plus 1 initial observation). A derivation of $T_{\min}$ and $T'_{\min}$ is available as supplementary material.

a new data set was generated.[20] When an SPJ estimator required a subpanel MLE that was indeterminate or infinite, it was replaced as follows: $\widehat{\theta}_{1/\{2,3\}}$ and $\widehat{\theta}_{1/\{3/2,2\}}$ by $\widehat{\theta}_{1/2}$, and $\widehat{\theta}_{1/2}$ by $\widehat{\theta}$.

Tables 3 and 4 report the biases and root mean-squared errors (RMSEs) of the estimators, along with the coverage rates of the bootstrap 95% confidence intervals ($\mathrm{CI}_{.95}$, based on 39 bootstrap draws).[21] In both models and at all design points, all bias-corrected estimators have less bias than the MLE. In most cases, the bias reduction is quite substantial. The asymptotic bias orders—which are $O(T^{-1})$ for $\widehat{\theta}$; $O(T^{-2})$ for $\widehat{\theta}_{1/2}$, $\dot{\theta}_{1/2}$, $\widehat{\theta}_{\mathrm{HK}}$, and $\dot{\theta}_{\mathrm{AH}}$; and $O(T^{-3})$ for $\widehat{\theta}_{1/\{2,3\}}$ and $\dot{\theta}_{1/\{2,3\}}$—appear most clearly for the MLE and the SPJ estimators, although the bias of $\widehat{\theta}_{1/2}$ vanishes slightly faster than predicted by the theory. For the analytical corrections $\widehat{\theta}_{\mathrm{HK}}$ and $\dot{\theta}_{\mathrm{AH}}$, in contrast, the bias decreases somewhat too slowly as T grows. This suggests that the choice of bandwidth required by $\widehat{\theta}_{\mathrm{HK}}$ and $\dot{\theta}_{\mathrm{AH}}$ (set at 1 here while it should grow with T) is of key importance. No estimator uniformly dominates all the others in terms of bias, although $\dot{\theta}_{1/\{2,3\}}$ in many cases has the least bias, closely followed by $\widehat{\theta}_{1/2}$. While the SPJ estimators are typically somewhat more variable than the MLE, this is more than offset by the removal of the leading bias terms, which results in much smaller RMSEs. This is in line with the discussion in Subsection 3.1. In our designs, $\widehat{\theta}_{1/\{2,3\}}$ often has a larger RMSE than $\widehat{\theta}_{1/2}$. The difference, however, decreases quickly as T increases. By contrast, $\dot{\theta}_{1/\{2,3\}}$ has a smaller RMSE than $\dot{\theta}_{1/2}$, uniformly over all designs, due to its success in reducing the bias. Interestingly, the analytically bias-corrected estimators generally have a smaller standard deviation (not reported) than does the MLE and, hence, than do the SPJ estimators. However, as they typically remove a smaller fraction of the bias, their RMSE is often larger than that of the SPJ estimators; the exception is $\widehat{\beta}_{\mathrm{HK}}$, which has little bias. Except in a few cases where the bias is substantial, the confidence intervals based on the SPJ estimators have reasonable coverage rates. For $\dot{\theta}_{1/2}$ and the analytical corrections, the ratio of bias to standard deviation is typically larger, so the coverage rates of their confidence intervals are worse. When T is not too small, however, they are still far better than those based on the MLE.

[20]From the point of view of drawing inference, this is unproblematic when $\widehat{\theta}$ is indeterminate because then the data are uninformative. However, discarding the data when $\widehat{\theta}$ is infinite is more problematic because then the data may in fact be quite informative, which calls for some other approach. The whole issue is probably empirically unimportant but has to be taken care of in simulations with small N and T.

[21]A small number of bootstrap draws suffices for studying coverage rates. In applications, a larger number should be used.

Table 3: Probit AR(1), common parameter, $N = 500$

T	ρ_0	$\widehat{\rho}$	$\widehat{\rho}_{1/2}$	$\widehat{\rho}_{1/\{2,3\}}$	$\widehat{\rho}_{\text{HK}}$	$\dot{\rho}_{1/2}$	$\dot{\rho}_{1/\{2,3\}}$	$\dot{\rho}_{\text{AH}}$
				bias				
6	.5	−.616	.228	−*.224*	−.247	−.270	−.068	−.278
9	.5	−.400	.055	−.107	−.135	−.128	−.012	−.155
12	.5	−.297	.021	−.027	−.091	−.071	−.002	−.104
18	.5	−.197	.006	−.005	−.054	−.032	.000	−.061
6	1	−.710	.152	−*.232*	−.389	−.402	−.199	−.414
9	1	−.471	.019	−.111	−.243	−.207	−.062	−.261
12	1	−.354	.001	−.027	−.178	−.120	−.021	−.190
18	1	−.238	−.003	−.006	−.114	−.056	−.005	−.122
				RMSE				
6	.5	.620	.251	*.312*	.255	.278	.116	.286
9	.5	.404	.088	.171	.145	.140	.079	.163
12	.5	.301	.059	.094	.101	.086	.064	.113
18	.5	.201	.042	.058	.064	.050	.048	.071
6	1	.741	.194	*.336*	.395	.409	.223	.419
9	1	.475	.080	.192	.249	.216	.104	.267
12	1	.358	.062	.110	.182	.131	.074	.196
18	1	.241	.046	.069	.120	.070	.055	.128
				$\text{CI}_{.95}$				
6	.5	.000	.491	*.836*	.052	.051	.893	.022
9	.5	.000	.892	.876	.305	.450	.945	.191
12	.5	.000	.937	.935	.513	.728	.948	.400
18	.5	.002	.951	.951	.720	.884	.951	.649
6	1	.000	.784	*.860*	.001	.002	.571	.000
9	1	.000	.945	.893	.013	.117	.894	.006
12	1	.000	.948	.939	.059	.444	.937	.036
18	1	.001	.951	.948	.210	.771	.950	.153

Italics: $\{3/2, 2\}$ instead of $\{2, 3\}$.

Table 4: Probit ARX(1), common parameters, $N = 500$

T	ρ_0	$\widehat{\rho}$	$\widehat{\rho}_{1/2}$	$\widehat{\rho}_{1/\{2,3\}}$	$\widehat{\rho}_{\text{HK}}$	$\dot{\rho}_{1/2}$	$\dot{\rho}_{1/\{2,3\}}$	$\dot{\rho}_{\text{AH}}$	$\widehat{\beta}$	$\widehat{\beta}_{1/2}$	$\widehat{\beta}_{1/\{2,3\}}$	$\widehat{\beta}_{\text{HK}}$	$\dot{\beta}_{1/2}$	$\dot{\beta}_{1/\{2,3\}}$	$\dot{\beta}_{\text{AH}}$
								bias							
6	.5	−.594	.279	*−.282*	−.224	−.261	−.073	−.270	.174	−.067	*.049*	−.019	.096	.050	.125
9	.5	−.385	.063	−.138	−.124	−.127	−.019	−.151	.112	−.020	.024	−.004	.050	.021	.063
12	.5	−.285	.024	−.033	−.083	−.070	−.005	−.101	.082	−.009	.006	.000	.030	.010	.039
18	.5	−.188	.009	−.004	−.048	−.031	.001	−.058	.053	−.005	−.001	−.002	.014	.003	.020
6	1	−.638	.207	*−.268*	−.350	−.352	−.170	−.356	.196	−.049	*.048*	−.063	.123	.076	.152
9	1	−.423	.031	−.134	−.218	−.184	−.056	−.228	.128	−.011	.025	−.020	.067	.035	.082
12	1	−.317	.005	−.036	−.156	−.107	−.020	−.166	.096	−.004	.008	−.006	.042	.018	.053
18	1	−.211	.001	−.003	−.098	−.047	−.001	−.105	.063	−.002	.000	−.001	.021	.006	.029
								RMSE							
6	.5	.599	.303	*.380*	.234	.271	.121	.279	.181	.099	*.200*	.038	.107	.074	.134
9	.5	.389	.095	.201	.134	.140	.079	.160	.117	.044	.097	.027	.060	.045	.071
12	.5	.289	.061	.100	.094	.086	.064	.111	.086	.031	.057	.022	.039	.033	.046
18	.5	.192	.042	.060	.060	.050	.049	.069	.057	.021	.033	.018	.024	.024	.028
6	1	.644	.245	*.393*	.357	.361	.199	.365	.204	.096	*.229*	.070	.134	.098	.161
9	1	.428	.084	.214	.224	.194	.100	.235	.134	.047	.113	.034	.077	.057	.090
12	1	.322	.062	.114	.163	.120	.073	.173	.100	.034	.067	.024	.051	.040	.060
18	1	.215	.046	.069	.105	.064	.054	.112	.067	.023	.039	.019	.030	.027	.036
								$\text{CI}_{.95}$							
6	.5	.000	.383	*.819*	.124	.102	.891	.069	.084	.851	*.942*	.915	.493	.853	.289
9	.5	.000	.878	.853	.402	.482	.945	.262	.131	.922	.942	.946	.697	.918	.530
12	.5	.000	.932	.934	.593	.742	.948	.457	.182	.938	.943	.947	.813	.940	.688
18	.5	.007	.943	.951	.763	.890	.950	.687	.299	.940	.951	.942	.895	.946	.828
6	1	.000	.690	*.862*	.006	.029	.683	.018	.086	.911	*.945*	.550	.405	.772	.226
9	1	.000	.939	.879	.048	.237	.906	.053	.122	.940	.939	.889	.584	.875	.417
12	1	.001	.948	.930	.138	.544	.936	.121	.148	.946	.949	.941	.727	.913	.564
18	1	.004	.947	.949	.342	.824	.947	.296	.239	.946	.950	.947	.856	.942	.744

Italics: $\{3/2, 2\}$ instead of $\{2, 3\}$.

As average effects on the probability that $y_{it} = 1$, consider

$$\mu_0 \equiv \overline{\mathbb{E}}\left(\Phi(\alpha_{i0} + \rho_0) - \Phi(\alpha_{i0})\right)$$

in the AR(1), with values $\mu_0 = .138, .260$ corresponding to $\rho_0 = .5, 1$, and

$$\mu_0^y \equiv \overline{\mathbb{E}}\left(\Phi(\alpha_{i0} + \rho_0 + \beta_0 x_{it}) - \Phi(\alpha_{i0} + \beta_0 x_{it})\right), \qquad \mu_0^x \equiv \beta_0 \overline{\mathbb{E}} \phi(\alpha_{i0} + \rho_0 + \beta_0 x_{it}),$$

in the ARX(1), with values $(\mu_0^y, \mu_0^x) = (.128, .124), (.244, .105)$ corresponding to $\rho_0 = .5, 1$. We estimated μ_0, μ_0^y, and μ_0^x by the corresponding sample average with the MLE serving as a plug-in estimate, for example,

$$\widehat{\mu}^y \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\Phi(\widehat{\alpha}_i + \widehat{\rho} + \widehat{\beta} x_{it}) - \Phi(\widehat{\alpha}_i + \widehat{\beta} x_{it}))$$

for μ_0^y, and the SPJ estimators, for example, $\widehat{\mu}_{1/2}^y$ and $\widehat{\mu}_{1/\{2,3\}}^y$ (or $\widehat{\mu}_{1/\{3/2,2\}}^y$) obtained by jackknifing $\widehat{\mu}^y$.

Tables 5 and 6 present the biases and RMSEs of the average effect estimators. Across the designs, the MLE-based estimates have large biases, exceeding 50% of the value of the estimand in more than half of the cases. The SPJ, especially the second-order SPJ, eliminates much of this bias. In addition, their RMSE is much lower. For μ_0 and μ_0^y, the second-order SPJ uniformly dominates the first-order SPJ in terms of bias and RMSE. For μ_0^x, the biases of the the SPJ estimators are very small, and the first-order SPJ has the lowest RMSE.

Table 5: Probit AR(1), average effect, $N = 500$

			bias			RMSE		
T	ρ_0	μ_0	$\widehat{\mu}$	$\widehat{\mu}_{1/2}$	$\widehat{\mu}_{1/\{2,3\}}$	$\widehat{\mu}$	$\widehat{\mu}_{1/2}$	$\widehat{\mu}_{1/\{2,3\}}$
6	.5	.138	−.159	−.086	*−.015*	.159	.088	*.041*
9	.5	.138	−.117	−.043	−.008	.118	.046	.025
12	.5	.138	−.092	−.025	−.003	.093	.029	.020
18	.5	.138	−.064	−.011	.000	.065	.017	.016
6	1	.260	−.218	−.137	*−.052*	.219	.138	*.065*
9	1	.260	−.167	−.079	−.034	.167	.081	.043
12	1	.260	−.134	−.050	−.017	.135	.053	.029
18	1	.260	−.096	−.025	−.005	.096	.029	.020

Italics: $\{3/2, 2\}$ instead of $\{2, 3\}$.

We provide the results of further Monte Carlo experiments as supplementary material. For the probit models, we considered the effect of decreasing N from 500 to

Table 6: Probit ARX(1), average effects, $N = 500$

			bias			RMSE		
T	ρ_0	μ_0^y	$\widehat{\mu}^y$	$\widehat{\mu}^y_{1/2}$	$\widehat{\mu}^y_{1/\{2,3\}}$	$\widehat{\mu}^y$	$\widehat{\mu}^y_{1/2}$	$\widehat{\mu}^y_{1/\{2,3\}}$
6	.5	.128	−.154	.079	*−.047*	.155	.085	*.080*
9	.5	.128	−.097	.025	−.024	.099	.032	.044
12	.5	.128	−.071	.011	−.008	.073	.019	.026
18	.5	.128	−.047	.004	-.002	.048	.012	.017
6	1	.244	−.143	.095	*−.045*	.145	.102	*.093*
9	1	.244	−.086	.037	−.015	.088	.044	.050
12	1	.244	−.060	.022	.003	.062	.029	.032
18	1	.244	−.036	.013	.006	.038	.020	.022
T	ρ_0	μ_0^x	$\widehat{\mu}^x$	$\widehat{\mu}^x_{1/2}$	$\widehat{\mu}^x_{1/\{2,3\}}$	$\widehat{\mu}^x$	$\widehat{\mu}^x_{1/2}$	$\widehat{\mu}^x_{1/\{2,3\}}$
6	.5	.124	.060	.003	*−.019*	.060	.016	*.046*
9	.5	.124	.039	−.003	−.009	.040	.009	.022
12	.5	.124	.029	−.003	−.003	.029	.007	.013
18	.5	.124	.019	−.002	−.002	.019	.005	.008
6	1	.105	.085	.010	*−.003*	.086	.020	*.049*
9	1	.105	.058	.005	.001	.059	.011	.023
12	1	.105	.045	.004	.003	.045	.009	.015
18	1	.105	.031	.003	.002	.031	.006	.009

Italics: $\{3/2, 2\}$ instead of $\{2, 3\}$.

100, and the effect of non-stationarity by setting $y_{i0} = 0$ for all i. As expected, setting $N = 100$ results in larger RMSE for all estimators; their biases, however, remain virtually unaltered. As a consequence, the coverage rates of all estimators improve. Regarding marginal effects, the second-order SPJ is still superior in terms of bias for most designs, but its RMSE is now often slightly larger than that of the first-order SPJ, especially in the ARX(1). When $y_{i0} = 0$, the overal pattern changes little, which suggests that our methods are reasonably robust with respect to mild violations of the stationarity requirement. We found that $\widehat{\theta}$, $\widehat{\theta}_{\text{HK}}$, and $\dot{\theta}_{\text{AH}}$ generally have a somewhat smaller bias and RMSE and slightly improved confidence intervals. For the SPJ estimates, the effect on the bias is mixed, the first-order SPJ often being slightly more biased and the second-order SPJ somewhat less. Their RMSEs follow the same pattern. The picture for the marginal effect estimates changes very little.

The supplementary material also contains simulation results for dynamic logit models using the same designs as for the probit models there. To facilitate comparison with the probit models, the errors in the logit models were normalized to have unit variance. The results are very similar to those for the probit models. Here too, the MLE is heavily biased and the SPJ is very effective at bias reduction, which leads to substantially smaller RMSEs and improved coverage rates for the common parameters. Likewise, the SPJ estimates of the average effects have less bias and a lower RMSE.

7 Conclusion

A split-panel jackknife estimator was derived for reducing the bias of the maximum likelihood estimator in nonlinear dynamic panel data models with fixed effects. The asymptotic distribution of the resulting estimates is normal and correctly centered under slow T asymptotics without inflating the asymptotic variance. The SPJ implicitly estimates the bias of the MLE up to the chosen order and, hence, can be viewed as an automatic bias-correction method. The SPJ is conceptually and computationally very simple as it requires only a few maximum likelihood estimates. There is no analytical work involved. We also gave jackknife corrections to the profile loglikelihood and discussed bias correction for average effects. The extension to other extremum estimators such as GMM is immediate, provided the asymptotic bias of the estimator or minimand admits an expansion in powers of T^{-1}.

Our results and subsequent recommendations are based on asymptotics where the number of time periods grows, fast or slowly, with the number of cross-sectional units. To refine those recommendations, more specifically about how to choose the order of bias reduction for given N and T, higher-order approximations to the variance of the MLE and the SPJ estimators would be of great interest. A related question is how the first-order SPJ relates to the analytical corrections of Hahn and Kuersteiner (2004) and Arellano and Hahn (2006) at the order $O(T^{-2})$. Another challenging issue is that of non-stationarity and, in particular, if and how the SPJ can be modified to accommodate the inclusion of time dummies or time trends. Allowing for such effects is important in a variety of microeconometric applications. Fixed T consistent estimators may break down in such a situation, possibly because of the loss of point identification. See, for example, Honoré and Kyriazidou's (2000) estimators for dynamic discrete-choice models and Honoré and Tamer (2006) on the lack of point identification in the presence of time effects.

In a simulation study of dynamic binary-choice models the SPJ was found to perform well even in short panels, showing much smaller biases and RMSEs than the MLE and confidence intervals with, mostly, acceptable coverage rates. It would be of interest to see how the split-panel jackknife and the various bias-corrected estimators proposed elsewhere perform in a broader range of models.

Appendix: Proofs

Proof of Theorem 1. Since $|S|/T$ is bounded away from zero for all $S \in \mathcal{S}$,

$$\text{plim}_{N\to\infty}\overline{\theta}_{\mathcal{S}} = \theta_0 + \sum_{j=1}^{k}\sum_{S\in\mathcal{S}}\frac{|S|^{1-j}}{T}B_j + o(T^{-k})$$
$$= \theta_0 + \frac{g}{T}B_1 + \sum_{j=2}^{k}\sum_{S\in\mathcal{S}}\frac{|S|^{1-j}}{T}B_j + o(T^{-k}),$$

where, by convention, $\sum_{j=2}^{k}(\cdot) = 0$ if $k = 1$. The result regarding $\text{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}}$ follows easily, since $B'_j = O(1)$ for $j = 2, ..., k$ because $T/|S| = O(1)$ for all $S \in \mathcal{S}$. Since $T/|S| > 1$, we have $T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j} > g$ for all $j \geq 2$, so $\text{sign}(B'_j) = -\text{sign}(B_j)$. To prove that $|B'_j| \geq |B_j|\sum_{m=1}^{j-1} g^m$, it suffices to show that, for $j \geq 2$,

$$T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j} - g \geq (g-1)\sum_{m=1}^{j-1} g^m. \tag{A.1}$$

By a property of the harmonic mean, for $j \geq 2$,

$$T^{j-1}\sum_{S\in\mathcal{S}}|S|^{1-j} \geq T^{j-1}\sum_{S\in\mathcal{S}}\left(\frac{T}{g}\right)^{1-j} = g^j,$$

from which (A.1) follows. As regards the asymptotic distribution of $\widehat{\theta}_{\mathcal{S}}$, note that, for any distinct $S, S' \in \mathcal{S}$, because S and S' are disjoint, $\sqrt{NT}(\widehat{\theta}_S - \theta_{|S|})$ and $\sqrt{NT}(\widehat{\theta}_{S'} - \theta_{|S'|})$ are jointly asymptotically normal as $N, T \to \infty$, with large N, T covariance equal to zero. Then, from Assumptions 1 and 3 and noting that $\mathcal{S}$ is a partition of $\{1, ..., T\}$, it follows that, as $N, T \to \infty$,

$$\sqrt{NT}\begin{pmatrix} \widehat{\theta} - \theta_T \\ \overline{\theta}_{\mathcal{S}} - \text{plim}_{N\to\infty}\overline{\theta}_{\mathcal{S}} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Omega & \Omega \\ \Omega & \Omega \end{pmatrix}\right)$$

and, in turn, $\sqrt{NT}(\widehat{\theta}_{\mathcal{S}} - \text{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}}) \xrightarrow{d} \mathcal{N}(0, \Omega)$. If, in addition, $N/T^3 \to 0$, then $\sqrt{NT}(\text{plim}_{N\to\infty}\widehat{\theta}_{\mathcal{S}} - \theta_0) = \sqrt{NT}O(T^{-2}) \to 0$ and (3.2) follows.

Proof of Theorem 2. For all $g \in G$,

$$\text{plim}_{N\to\infty}\overline{\theta}_{1/g} = \theta_0 + \sum_{j=1}^{k}\sum_{S\in\mathcal{S}_g}\frac{|S|^{1-j}}{T}B_j + o(T^{-k}).$$

Hence

$$\operatorname{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + \sum_{j=1}^{k}\left(\left(1+\sum_{g\in G}a_{1/g}\right)\frac{1}{T^j} - \sum_{g\in G}a_{1/g}\sum_{S\in\mathcal{S}_g}\frac{|S|^{1-j}}{T}\right)B_j + o(T^{-k})$$
$$= \theta_0 + \sum_{j=1}^{k}\frac{c_j(G)B_j}{T^j} + o(T^{-k}),$$

where

$$c_j(G) \equiv 1 + \sum_{g\in G}a_{1/g}\left(1 - \sum_{S\in\mathcal{S}_g}T^{j-1}|S|^{1-j}\right)$$
$$= (1-\iota' A^{-1}\iota)^{-1} - \sum_{g\in G}a_{1/g}\sum_{S\in\mathcal{S}_g}T^{j-1}|S|^{1-j}$$
$$= (1-\iota' A^{-1}\iota)^{-1} - \sum_{r=1}^{l}a_{1/g_r}\sum_{S\in\mathcal{S}_{g_r}}T^{j-1}|S|^{1-j}$$
$$= (1-\iota' A^{-1}\iota)^{-1}\left(1 - \sum_{r=1}^{l}\left(\sum_{s=1}^{l}A^{rs}\right)\sum_{S\in\mathcal{S}_{g_r}}T^{j-1}|S|^{1-j}\right), \qquad \text{(A.2)}$$

and A^{rs} is the (r,s)th element of A^{-1}. For $j \leq l$,

$$c_j(G) = (1-\iota' A^{-1}\iota)^{-1}\left(1 - \sum_{r=1}^{l}\left(\sum_{s=1}^{l}A^{rs}\right)A_{jr}\right)$$
$$= (1-\iota' A^{-1}\iota)^{-1}\left(1 - \sum_{s=1}^{l}\sum_{r=1}^{l}A_{jr}A^{rs}\right) = 0.$$

This proves that $\operatorname{plim}_{N\to\infty}\widehat{\theta}_{1/G} = \theta_0 + o\left(T^{-l}\right)$ if $k = l$. Now consider the case $k > l$. We need to show that $c_j(G) = b_j(G) + O(T^{-1})$ for $l < j \leq k$. For all $g \in G$ and all $S \in \mathcal{S}_g$, $T|S|^{-1} = g + O(T^{-1})$, and, for $r = 1, ..., k$, $\sum_{S\in\mathcal{S}_g}T^{r-1}|S|^{1-r} = g^r + O(T^{-1})$. Hence $A = \mathbf{A} + O(T^{-1})$, where $\mathbf{A}$ is the $l \times l$ matrix with elements $\mathbf{A}_{rs} = g_s^r$. Let

$\pi_j \equiv (g_1^j, ..., g_l^j)'$. From (A.2), for $l < j \leq k$,

$$
\begin{aligned}
c_j(G) &= (1 - \iota'\mathbf{A}^{-1}\iota)^{-1} \left(1 - \sum_{r=1}^{l} \left(\sum_{s=1}^{l} \mathbf{A}^{rs} \right) g_r^j \right) + O(T^{-1}) \\
&= (1 - \iota'\mathbf{A}^{-1}\iota)^{-1} \left(1 - \pi_j'\mathbf{A}^{-1}\iota \right) + O(T^{-1}) \\
&= \frac{|\mathbf{A}|^{-1} \begin{vmatrix} \mathbf{A} & \iota \\ \pi_j' & 1 \end{vmatrix}}{|\mathbf{A}|^{-1} \begin{vmatrix} \mathbf{A} & \iota \\ \iota' & 1 \end{vmatrix}} + O(T^{-1}) = (-1)^l \frac{|V_j|}{|V|} + O(T^{-1}),
\end{aligned}
$$

where ι is an $l \times 1$ vector of ones and

$$
|V| = \begin{vmatrix} 1 & \iota' \\ \iota & \mathbf{A}' \end{vmatrix}, \qquad V_j = \begin{vmatrix} \iota' & 1 \\ \mathbf{A}' & \pi_j \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 1 & \iota' & 1 \\ \iota & \mathbf{A}' & \pi_j \end{vmatrix}.
$$

$|V|$ is a Vandermonde determinant given by

$$
|V| = \prod_{0 \leq p < q \leq l} (g_q - g_p), \qquad g_0 \equiv 1.
$$

Noting that the first row of V_{l+1} is $(0^0, 0^1, ..., 0^{l+1})$, $|V_{l+1}|$ is also a Vandermonde determinant, given by

$$
|V_{l+1}| = \prod_{-1 \leq p < q \leq l} (g_q - g_p) = |V| \prod_{1 \leq q \leq l} g_q, \qquad g_{-1} \equiv 0.
$$

For $j > l+1$, by the Jacobi-Trudi identity (see, e.g., Littlewood, 1958, pp. 88), $|V_j|$ can be written as the product of $|V_{l+1}|$ and a homogeneous product sum of $g_{-1}, g_0, ..., g_l$,

$$
|V_j| = |V_{l+1}| \sum_{\substack{k_{-1}, k_0, ..., k_l \geq 0 \\ k_{-1} + k_0 + ... + k_l = j - l - 1}} g_{-1}^{k_{-1}} g_0^{k_0} ... g_l^{k_l} = |V_{l+1}| \sum_{\substack{k_1, ..., k_l \geq 0 \\ k_1 + ... + k_l \leq j - l - 1}} g_1^{k_1} ... g_l^{k_l},
$$

which also holds for $j = l + 1$. On collecting results, $c_j(G) = b_j(G) + O(T^{-1})$ for $l < j \leq k$. The asymptotic distribution of $\widehat{\theta}_{1/G}$, under the asymptotics considered, follows along the lines of the proof of Theorem 1.

Proof of Theorem 3. The first part is proved along the same lines as in Theorem 2. We have

$$
\text{plim}_{N \to \infty} \widehat{\theta}_{1/G} = \theta_0 + \sum_{j=1}^{k} \frac{c_j(G) B_j}{T^j} + o(T^{-k}),
$$

where now

$$
\begin{aligned}
c_j(G) &\equiv 1+\sum_{g\in G} a_{1/g}\left(1-\sum_{S\in\mathcal{S}_g}\frac{T^j|S|^{1-j}}{\sum_{S\in\mathcal{S}_g}|S|}\right)\\
&=(1-\iota' A^{-1}\iota)^{-1}\left(1-\sum_{r=1}^{l}\left(\sum_{s=1}^{l}A^{rs}\right)\sum_{S\in\mathcal{S}_{g_r}}\frac{T^j|S|^{1-j}}{\sum_{S\in\mathcal{S}_{g_r}}|S|}\right).
\end{aligned}
$$

For $j\le l$, $c_j(G)=0$. Consider the case $k>l$. For all $g\in G$ and $r=1,...,k$,

$$
\begin{aligned}
\sum_{S\in\mathcal{S}_g}\frac{T^r|S|^{1-r}}{\sum_{S\in\mathcal{S}_g}|S|} &= \frac{T}{\sum_{S\in\mathcal{S}_g}|S|}\sum_{S\in\mathcal{S}_g}T^{r-1}|S|^{1-r} = \frac{g}{\sum_{S\in\mathcal{S}_g}1}g^{r-1}\sum_{S\in\mathcal{S}_g}1+O(T^{-1})\\
&= g^r+O(T^{-1}).
\end{aligned}
$$

Hence, $A=\mathbf{A}+O(T^{-1})$ and, for $l<j\le k$,

$$
c_j(G)=(1-\iota'\mathbf{A}^{-1}\iota)^{-1}\left(1-\pi_j'\mathbf{A}^{-1}\iota\right)+O(T^{-1}),
$$

where $\pi_j\equiv(g_1^j,...,g_l^j)'$. By the proof of Theorem 2, $c_j(G)=b_j(G)+O(T^{-1})$ for $l<j\le k$, thus completing the proof of the first part. We now derive the asymptotic distribution of $\widehat{\theta}_{1/G}$. For any subpanels S and S' such that, as $T\to\infty$, $T^{-1}|S|\to s>0$, $T^{-1}|S'|\to s'>0$, and $T^{-1}|S\cap S'|\to s_\cap\ge 0$, we have

$$
\text{Avar}\begin{pmatrix}\widehat{\theta}_S\\ \widehat{\theta}_{S'}\end{pmatrix}=\begin{pmatrix}1/s & s_\cap/(ss')\\ s_\cap/(ss') & 1/s'\end{pmatrix}\otimes\Omega, \tag{A.3}
$$

where $\text{Avar}(\cdot)$ denotes the large N,T variance. Now consider $\overline{\theta}_{1/g}=\frac{1}{2}(\widehat{\theta}_{S_1}+\widehat{\theta}_{S_2})$ and $\overline{\theta}_{1/g'}=\frac{1}{2}(\widehat{\theta}_{S_1'}+\widehat{\theta}_{S_2'})$, where $1<g<g'<2$ and $1\in S_1\cap S_1'$. Then $T^{-1}|S_1|=T^{-1}|S_2|\to 1/g$, $T^{-1}|S_1'|=T^{-1}|S_2'|\to 1/g'$, $T^{-1}|S_1\cap S_2|\to(2-g)/g$, $T^{-1}|S_1'\cap S_2'|\to(2-g')/g'$, $T^{-1}|S_1\cap S_1'|=T^{-1}|S_2\cap S_2'|\to 1/g'$, and $T^{-1}|S_1\cap S_2'|=T^{-1}|S_2\cap S_1'|\to(g+g'-gg')/(gg')$. Application of (A.3) gives

$$
\text{Avar}\begin{pmatrix}\widehat{\theta}_{S_1}\\ \widehat{\theta}_{S_2}\\ \widehat{\theta}_{S_1'}\\ \widehat{\theta}_{S_2'}\end{pmatrix}=\begin{pmatrix} g & g(2-g) & g & g+g'-gg'\\ g(2-g) & g & g+g'-gg' & g\\ g & g+g'-gg' & g' & g'(2-g')\\ g+g'-gg' & g & g'(2-g') & g'\end{pmatrix}\otimes\Omega,
$$

and so

$$
\text{Avar}\begin{pmatrix}\overline{\theta}_{1/g}\\ \overline{\theta}_{1/g'}\end{pmatrix}=\frac{1}{2}\begin{pmatrix} g(3-g) & 2g+g'-gg'\\ 2g+g'-gg' & g'(3-g')\end{pmatrix}\otimes\Omega.
$$

Let $\overline{\theta}_{1/G} \equiv (\overline{\theta}_{1/g_1}, ..., \overline{\theta}_{1/g_l})$. Then $\text{Avar}(\text{vec}\,\overline{\theta}_{1/G}) = V \otimes \Omega$, where $\text{vec}(\cdot)$ is the stack operator and V is the symmetric $l \times l$ matrix whose $(r,s)^{\text{th}}$ element, for $r \leq s$, is

$$V_{rs} \equiv \begin{cases} g_r + \frac{1}{2}(g_s - g_r g_s) & \text{if } s \leq o, \\ 1 & \text{otherwise.} \end{cases}$$

Therefore, $\widehat{\theta}_{1/G} = (1 - \iota' A^{-1}\iota)^{-1}(\widehat{\theta} - \overline{\theta}_{1/G} A^{-1}\iota)$ is asymptotically normally distributed, centered at θ_0, and has large N,T variance

$$\begin{aligned} \text{Avar}(\widehat{\theta}_{1/G}) &= (1 - \iota'\mathbf{A}^{-1}\iota)^{-2}(1 - 2\iota'\mathbf{A}^{-1}\iota + \iota'\mathbf{A}'^{-1}V\mathbf{A}^{-1}\iota)\Omega \\ &= \left(1 + \frac{\iota'\mathbf{A}'^{-1}\left(V - \iota\iota'\right)\mathbf{A}^{-1}\iota}{(1 - \iota'\mathbf{A}^{-1}\iota)^2}\right)\Omega = d(G)\Omega, \end{aligned}$$

since $V - \iota\iota' = \Gamma$. The proof is completed by showing that, if $o \geq 1$, the leading $o \times o$ submatrix of Γ is positive definite. Let L_o be 2 times this submatrix, so that

$$L_o = \begin{pmatrix} L_{o-1} & \lambda_{o-1} \\ \lambda'_{o-1} & \lambda_{oo} \end{pmatrix},$$

where

$$\lambda_{o-1} \equiv \begin{pmatrix} g_1 - 1 \\ \vdots \\ g_{o-1} - 1 \end{pmatrix} (2 - g_o), \qquad \lambda_{oo} \equiv (g_o - 1)(2 - g_o).$$

The (r,s)-th element of L_{o-1}^{-1}, for $r \leq s$, is

$$L_{o-1}^{rs} = \begin{cases} \frac{g_{r+1} - g_{r-1}}{(g_r - g_{r-1})(g_{r+1} - g_r)} & \text{if } r = s < o-1, \\ \frac{2 - g_{o-2}}{(g_{o-1} - g_{o-2})(2 - g_{o-1})} & \text{if } r = s = o-1, \\ -\frac{1}{g_{r+1} - g_r} & \text{if } r = s-1, \\ 0 & \text{otherwise,} \end{cases}$$

where $g_0 \equiv 1$. Hence

$$\lambda'_{o-1} L_{o-1}^{-1} \lambda_{o-1} = (2 - g_o)^2 \left(\sum_{r=1}^{o-2} h_r + \frac{(g_{o-1} - 1)^2 (2 - g_{o-2})}{(g_{o-1} - g_{o-2})(2 - g_{o-1})} \right),$$

where

$$\begin{aligned} h_r &= \frac{(g_r - 1)^2 (g_{r+1} - g_{r-1})}{(g_r - g_{r-1})(g_{r+1} - g_r)} - 2\frac{(g_r - 1)(g_{r+1} - 1)}{g_{r+1} - g_r} \\ &= (g_r - 1)\left(\frac{g_{r-1} - 1}{g_r - g_{r-1}} - \frac{g_{r+1} - 1}{g_{r+1} - g_r}\right). \end{aligned}$$

After some algebra, $\sum_{r=1}^{o-2} h_r = -\frac{(g_{o-1}-1)(g_{o-2}-1)}{g_{o-1}-g_{o-2}}$, and so

$$\lambda_{oo} - \lambda_{o-1}' L_{o-1}^{-1} \lambda_{o-1} = \frac{(g_o - g_{o-1})(2 - g_o)}{2 - g_{o-1}}.$$

The determinant of L_o is

$$|L_o| = |L_{o-1}| \left(\lambda_{oo} - \lambda_{o-1}' L_{o-1}^{-1} \lambda_{o-1} \right) = (2 - g_o) \prod_{r=1}^{o} (g_r - g_{r-1}),$$

by induction. Clearly, $0 < |L_o| < |L_{o-1}| < ... < |L_1| < 1$. All leading submatrices of L_o have a positive determinant, so L_o is positive definite.

References

[1] Adams, J., H. Gray, and T. Watkins. (1971). An asymptotic characterization of bias reduction by jackknifing. *Annals of Mathematical Statistics*, 42:1606–1612.

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

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

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

[5] Anderson, T. W. and C. Hsiao. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18:47–82, 1982.

[6] Arellano, M. (2003). Discrete choices with panel data. *Investigaciones Económicas*, 27:423–458.

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

[8] Arellano, M. and J. Hahn. (2006). A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects. Unpublished manuscript.

[9] Arellano, M. and J. Hahn. (2007). Understanding bias in nonlinear panel models: Some recent developments. In R. W. Blundell, W. K. Newey, and P. Torsten (eds.), *Advances In Economics and Econometrics*, Volume III. Econometric Society, Cambridge University Press.

[10] Arellano, M. and B. E. Honoré. (2001). Panel data models: Some recent developments. In J. J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Volume 5, Chapter 53, pp. 3229–3329. Elsevier.

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

[12] Brillinger, D. R. (1964). The asymptotic behavior of Tukey's general method of setting approximate confidence-limits (the jackknife) when applied to maximum likelihood estimates. *Review of the Institute of International Statistics*, 32:202–206.

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

[14] Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies* , 47:225–238.

[15] Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, Volume 2, Chapter 22, pp. 1247–1315. Elsevier.

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

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

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

[19] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38.

[20] Greene, W. H. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *Econometrics Journal*, 7:98–119.

[21] Hahn, J. (2001). The information bound of a dynamic panel logit model with fixed effects. *Econometric Theory*, 17:913–932.

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

[23] Hahn, J. and G. Kuersteiner. (2004). Bias reduction for dynamic nonlinear panel models with fixed effects. Unpublished manuscript.

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

[25] Hall, B. H. (1978). A general framework for the time series-cross section estimation. *Annales de L'INSEE*, 30/31:177–202.

[26] Honoré, B. E. and E. Kyriazidou. (2002). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68:839–874.

[27] Honoré, B. E. and E. Tamer. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, 74:611–629.

[28] Hu, L. (2002). Estimation of a censored dynamic panel data model. *Econometrica*, 70:2499–2517.

[29] Kapetanios, G. (2008). A bootstrap procedure for panel data sets with many cross-sectional units. *Econometrics Journal*, 11:377–395.

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

[31] Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies*, 69:647–666.

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

[33] Littlewood, D. E. (1958). *The Theory of Group Characters and Matrix Representations of Groups*. Oxford, Clarendon, second edition.

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

[35] Miller, R. G. (1974). The jackknife – a review. *Biometrika*, 61:1–15.

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

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

[38] Quenouille, H. M. (1949) Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11:68–84.

[39] Quenouille, H. M. (1956). Notes on bias in estimation. *Biometrika*, 43:353–360.

[40] Rasch, G. (1961). On the general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4. University of California Press, Berkeley and Los Angeles.

[41] Reeds, J. A. (1978). Jackknifing maximum likelihood estimates. *Annals of Statistics*, 6:727–739.

[42] Sartori, N. (2003). Modified profile likelihood in models with stratum nuisance parameters. *Biometrika*, 90:533–549.

[43] Schucany, W. R., H. Gray, and D. Owen. (1971). On bias reduction in estimation. *Journal of the American Statistical Association*, 66:524–533.

[44] Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford Statistical Science Series. Oxford University Press.

[45] Shao, J. and D. Tu. (1995). *The Jackknife and Bootstrap*. Springer-Verlag.

[46] Sweeting, T. J. (1987). Discussion of the paper by Professors Cox and Reid. *Journal of the Royal Statistical Society, Series B*, 49:20–21.

[47] Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*, 29:614.

[48] Woutersen, T. (2002) Robustness against incidental parameters. Unpublished manuscript.